



ROUTLEDGE
INTERNATIONAL
HANDBOOKS



The Routledge International Handbook of Advanced Quantitative Methods in Nursing Research

Edited by Susan J. Henly

ROUTLEDGE INTERNATIONAL HANDBOOK OF ADVANCED QUANTITATIVE METHODS IN NURSING RESEARCH

Designed to support global development of nursing science, the *Routledge International Handbook of Advanced Quantitative Methods in Nursing Research* provides a new, comprehensive, and authoritative treatment of advanced quantitative methods for nursing research.

Incorporating past approaches that have served as the foundation for the science, this cutting-edge book also explores emerging approaches that will shape its future. Divided into six parts, it covers:

- the domain of nursing science;
- measurement – classical test theory, IRT, clinimetrics, behavioral observation, biophysical measurement;
- models for prediction and explanation – SEM, general growth mixture models, hierarchical models, analysis of dynamic systems;
- intervention research – theory-based interventions, causality, third variables, pilot studies, quasi-experimental design, joint models for longitudinal data and time to event;
- e-science – DIKW paradigm, big data, data mining, omics, fMRI;
- special topics – comparative effectiveness and meta-analysis, patient safety, economics research in nursing, mixed methods, global research dissemination.

Written by a distinguished group of international nursing scientists, scientists from related fields, and methodologists, the *Handbook* is the ideal reference for everyone involved in nursing science, whether they are graduate students, academics, editors and reviewers, or clinical investigators.

Susan J. Henly is Professor Emerita, University of Minnesota, School of Nursing, Minneapolis, Minnesota, USA and Editor of *Nursing Research*. She earned her BS with a major in nursing from the College of St. Teresa, Winona, Minnesota and practiced in rural acute care, perinatal nursing, and neonatal intensive care in Alaska, New Mexico, and Minnesota before returning to graduate school. Her MS degree in nursing, focused on perinatal health research, is from the University of Minnesota, Twin Cities. She earned her PhD in psychometric methods from the University of Minnesota, Twin Cities. She served on the College of Nursing faculty at the University of North Dakota, Grand Forks, prior to her appointment at Minnesota.

Over the past 30 years, Sue's research has focused on psychometric methods for nursing research with special interests in robustness of estimators in the analysis of covariance structures, model selection, and longitudinal models for health trajectories. She was Methods Director for the National Institute of Nursing Research-Funded Center for Health Trajectory Research at the University of Minnesota, School of Nursing. Sue has a special interest in advancing quantitative methods in nursing PhD programs. She was director of the American Indian MS to PhD Nursing Science Bridge Program (funded by the National Institute of General Medical Sciences) and chaired the Council for the Advancement of Nursing Sciences Idea Festival for Nursing Science Education. She has extensive service as a peer reviewer for nursing science, related fields, and methodology journals and has contributed to the peer review literature. Sue is Editor of *Nursing Research*. She is a member of the Japan Academy of Nursing Science and the American Academy of Nursing.

ROUTLEDGE INTERNATIONAL HANDBOOK OF ADVANCED QUANTITATIVE METHODS IN NURSING RESEARCH

Edited by Susan J. Henly

First published 2016
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN
and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2016 S. J. Henly

The right of Susan J. Henly to be identified as the author of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

Routledge international handbook of advanced quantitative methods in nursing research /
edited by Susan J. Henly.
p. ; cm.

International handbook of advanced quantitative methods in nursing research
Includes bibliographical references and index.

I. Henly, Susan J. (Susan Joan), editor. II. Title: International handbook of advanced
quantitative methods in nursing research. [DNLM: 1. Nursing
Research—methods. 2. Research Design. WY 20.5]

RT81.5
610.73072—dc23
2015004365

ISBN: 978-0-415-52180-2 (hbk)
ISBN: 978-1-315-88230-7 (ebk)

Typeset in Bembo
by Apex CoVantage, LLC

To
My husband George
Our children Chris, Sam, and Johanna
My Dad

In Memory of
My Mother
Hank and Dot

This page intentionally left blank

CONTENTS

<i>List of figures</i>	<i>x</i>
<i>List of tables</i>	<i>xiii</i>
<i>Contributors</i>	<i>xv</i>
<i>Acknowledgments</i>	<i>xxi</i>
<i>Foreword</i>	<i>xxiii</i>
<i>Preface</i>	<i>xxv</i>

PART I

The domain of nursing science	1

1	The domain of nursing science	3
	<i>Patricia A. Grady and Jessica M. McIlvane</i>	
2	Theorizing in nursing science	15
	<i>Susan J. Henly</i>	

PART II

Measurement	27

3	Classical test theory	29
	<i>Klaas Sijtsma</i>	
4	Clinimetrics	44
	<i>Henrica de Vet and Anna Beurskens</i>	

5	Item response theory: a statistical theory of measurement based on fungible items	58
	<i>Peter Baldwin and Howard Wainer</i>	
6	Behavioral observation	81
	<i>Tondi M. Harrison</i>	
7	Biophysical measurement	98
	<i>Ulf G. Bronas and Diane Treat-Jacobson</i>	
PART III		
	Prediction and explanation	117
8	Structural equation modeling	119
	<i>Sarah J. Schmiege and Angela D. Bryan</i>	
9	General growth mixture models	137
	<i>Sarah A. Stoddard</i>	
10	Multilevel models	154
	<i>Ulrike Grittner and Nils Lahmann</i>	
11	Analysis of dynamical systems: the modeling of change and variability	170
	<i>Pascal R. Deboeck and Steven M. Boker</i>	
PART IV		
	Experimental and quasi-experimental design	187
12	Theory-based nursing interventions	189
	<i>Souraya Sidani and Julie Fleury</i>	
13	Pilot studies for randomized clinical trials	199
	<i>Nancy Feeley and Sylvie Cossette</i>	
14	Causality in experiments and observational studies	213
	<i>Donald B. Rubin and Elizabeth R. Zell</i>	
15	Quasi-experimental design in nursing research	224
	<i>Patricia Eckardt and David Rindskopf</i>	
16	Third variables: scientific meanings and modeling in non-randomized studies	238
	<i>Donna L. Coffman</i>	

17	Joint models for longitudinal data and time-to-event occurrence <i>Yue Liu and Lei Liu</i>	253
PART V		
	e-Science methods	265
18	Data, information, knowledge, wisdom <i>Nancy Staggers and Ramona Nelson</i>	267
19	Big data in nursing research <i>Bonnie L. Westra, Karen A. Monsen, and Connie W. Delaney</i>	280
20	Data mining and data visualization <i>Krzysztof J. Cios and Dat T. Nguyen</i>	294
21	Genomic, transcriptomic, epigenomic, and proteomic approaches <i>Yvette P. Conley</i>	324
22	A survey of the sources of noise in fMRI <i>Douglas N. Greve, Gregory G. Brown, Bryon A. Mueller, Gary H. Glover, and Thomas T. Liu</i>	336
PART VI		
	Special topics	359
23	Comparative effectiveness research and meta-analysis <i>Nicky Cullum and Jo Dumville</i>	361
24	Patient safety research: methodological challenges <i>Sean P. Clarke and Maria Schubert</i>	381
25	Economic evaluations for nursing research <i>Mayuko Uchida-Nakakoji and Patricia W. Stone</i>	397
26	Mixed methods <i>Louise Doyle</i>	411
27	Global generation and dissemination of nursing science <i>Molly C. Dougherty and Nalini Jairath</i>	423
	<i>Index</i>	433

FIGURES

2.1	Concept map: theorizing in nursing science	16
3.1	Two propensity distributions with different means (true scores) and standard deviations	30
4.1	An ordinal scale	48
4.2	The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN)	48
4.3	A Bland-Altman plot	51
4.4	Minimally important change and smallest detectable change	53
5.1	Item characteristic curves for two 1-PL items	61
5.2	Item characteristic curves for three 2-PL items with equal difficulties	62
5.3	Item characteristic curve for a 3-PL item	63
5.4	Item response category characteristic curves (IRCCCs) for a 5-category item	64
5.5	Boundary or operating characteristic functions for a 5-category item	65
5.6	The likelihood of responding positively to item 1 and negatively to item 2 as the product of each item's response function	68
5.7	Item information functions	73
6.1	Three initial steps in designing behavioral observation studies	84
6.2	Process of observing multiple participants using momentary time sampling by one observer	89
6.3	Sample data collection sheet with behavioral codes	90
7.1	Ultrasound wave forms	106
7.2	Forces on a solid object	109
7.3	Laminar flow of blood	110
7.4	Turbulent flow of arterial blood at a point of bifurcation	111
8.1	Example of a one-factor confirmatory factor analysis model	120
8.2	Example of an observed variable path model based on the theory of planned behavior	120
8.3	Example of a structural equation model with latent variables	121
9.1	General growth mixture models: theoretical hope/hopelessness trajectory model	138
9.2	Scatterplots: individual hopelessness measurements over time	145

9.3	Scatterplots: ordinary least squares intercepts and slopes	146
9.4	<i>Mplus</i> syntax: general growth mixture model	148
9.5	General growth mixture model: path model	149
9.6	Estimated hopelessness trajectories	151
10.1	Example of multilevel data structure	159
10.2	Regression of transformed Braden scores on age using three different models	162
11.1	Variability around growth trajectories	171
11.2	Derivatives for a plot of hypothetical depression scores across time	173
11.3	Linear model for the first derivative	176
11.4	Effects of a concurrently measured predictor on an outcome using a simulated first-order differential equation model with normally distributed errors	179
11.5	Moderating effects of z on β in a first-order differential equation model	180
13.1	Decision-making process from pilot to full-scale RCT	208
16.1	A moderator effect	239
16.2	Time-varying moderator effects	240
16.3	Confounding effects	241
16.4	Selection bias	241
16.5	A basic mediation model	242
16.6	Mediation models incorporating a time-varying post-treatment covariate	245
16.7	Instrumental variable effect, with confounder effect also shown	247
17.1	Exploratory plots of CD4 repeated measures	257
17.2	Residual plots	259
18.1	The Nelson data to wisdom model	268
19.1	A high-level structure for the Nursing Interventions Classification	281
19.2	Omaha System ontology	282
20.1	The six-step knowledge discovery process (KDP)	296
20.2	Phases of the feature selection process	299
20.3	Training and test example data sets	304
20.4	Training and test example for multiple-instance learning	305
20.5	Data for a one-class learning problem	306
20.6	Pseudocode of DataSqueezer algorithm	307
20.7	Toy example for the DataSqueezer algorithm step 1: Data Reduction	308
20.8	DataSqueezer algorithm, step 2: MakeRules	309
20.9	Illustration of the similarity between two bags	310
20.10	Pseudocode for the multiple instance data squeezer (mi-DS) algorithm	311
20.11	Determining the label for a new bag (patient) B	311
20.12	Rule generation using the mi-DS algorithm	312
20.13	Construction of a similarity matrix	313
20.14	New bag (patient) B data	313
20.15	Prediction of the label for new patient/bag B	313
20.16	Choosing an optimally complex model with acceptable fit and prediction errors	314
20.17	Illustration of the process for 10-fold cross-validation	315
20.18	Receiver operating curve (ROC) comparing performance of classifiers A and B	317
20.19	Threshold impact on misclassification	318
20.20	Pseudocode of the self-organizing map (SOM) algorithm	319

Figures

20.21	Illustration of interacting particles/points in 2-D space	320
20.22	Pseudocode for the molecular dynamics algorithm	320
20.23	Visualization of breast cancer data	321
22.1	Temporal waveforms used in analysis of the working memory task	338
22.2	Functional and anatomical images of the same subject	338
22.3	Raw and best fit time courses for a single voxel over a single run	339
22.4	Significance SPMs for the Probe-vs.-Scrambled Contrast	341
22.5	Residual noise power spectrum averaged over gray matter	352
23.1	Primary research participants, studies, and systematic reviews	362
23.2	Forest plot	373
23.3	Network of evidence	375
25.1	Sample decision tree	402
25.2	Basic steps in approaching economic evaluations	404

TABLES

1.1	The domain of nursing science	5
3.1	Example of a covariance matrix for four items	33
3.2	Two covariance matrices and corresponding correlation matrices producing the same alpha	35
4.1	Instrument databases	54
5.1	Boundary characteristic functions for 5-category GRM items	65
5.2	Some well-known commercial software packages for estimating IRT models	74
6.1	Studies using behavioral observation	82
6.2	Differentiating molar concepts and molecular behaviors	85
6.3	Time sampling methods	88
6.4	Threats to observational accuracy	94
7.1	Temperature scales	102
7.2	Bodily sounds assessed through auscultation with a stethoscope	105
7.3	Examples: research employing therapeutic use of sound or noise reduction in clinical care	106
7.4	Examples: electrical devices used for therapeutic interventions	108
7.5	Examples: interventions to enhance flow of bodily fluids	112
8.1	SEM matrix notation (following LISREL)	121
8.2	Features of commonly used SEM software packages	132
9.1	Initial data analysis: OLS linear function for change	147
9.2	Statistical fit for proposed multiple group models	148
9.3	Parameter estimates and class differences weighted by estimated class probabilities for final model	150
9.4	Estimated odds ratios for violence with a weapon during later adolescence by latent hopelessness class	151
10.1	Descriptive data for 10 selected hospitals	156
10.2	Multilevel linear model for transformed braden scores: random intercept model without covariates	160
10.3	Multilevel linear model for transformed braden scores: random intercept model with individual level covariate	161

10.4	Multilevel linear model for transformed braden scores: random intercepts and slopes model with individual-level covariate (10 hospitals)	163
10.5	Multilevel linear models for transformed braden scores: random intercepts and slopes models with individual-level covariates (22 hospitals)	165
13.1	Design elements and procedures: assessment for feasibility in a pilot RCT	202
13.2	Design elements and procedures: assessment for acceptability in a pilot RCT	204
15.1	Factors jeopardizing internal validity	227
15.2	Design variation exemplars	228
15.3	Single group quasi-experimental designs	229
15.4	Nonrandomized designs with a control group and no pretest or unconventional pretest design	230
15.5	Nonrandomized designs with both control groups and pretests	231
16.1	Definitions of causal effects	243
17.1	Joint analysis of CD4 count and survival	258
19.1	Comparison of elements: nursing minimum data set and nursing management minimum data set	283
19.2	Nursing minimum data sets, terminologies, and billing codes	284
19.3	Nursing informatics agenda for 2008–2018	290
20.1	Evaluation functions and feature selection algorithms	299
20.2	Quanta matrix	301
20.3	Concepts used in MIL	305
20.4	Misclassification matrix	316
20.5	Measures of classifier correctness	316
21.1	Internet resources for genomic, transcriptomic, epigenomic, and proteomic approaches	325
22.1	Sources of first level noise in fMRI	346
22.2	Sources of within subject/cross-visit noise in fMRI	349
22.3	Sources of cross-subject and cross-group level noise in fMRI	350
22.4	Sources of variation across sites in fMRI	351
23.1	Study eligibility considerations: protocol for systematic review	366
23.2	Systematic review: implementing protocol steps	369
23.3	Effect size measures for meta-analysis	371
24.1	Errors, mistakes, adverse events, and violations	383
24.2	Patient safety systems interventions	385
24.3	Culture of safety and leadership interventions	386
24.4	Limitations of data sources for safety research	387
24.5	Themes in patient safety research findings	392
25.1	Types of economic evaluations	398
25.2	Common terminology in economic evaluations	399
25.3	Basic cost elements of economic evaluations	405
25.4	CHEERS checklist	408

CONTRIBUTORS

Peter Baldwin is Senior Measurement Scientist at the National Board of Medical Examiners, Philadelphia, Pennsylvania, USA. His primary research interests include psychometric methods.

Anna J. Beurskens PhD is Professor of goal-oriented measurement in patient care at Maastricht University and program director of the research center Autonomy and Participation of Persons with a Chronic Illness at Zuyd University of Applied Sciences, Heerlen, the Netherlands. The central themes in her research and education projects are: using measurements in daily care, patient-centered goal setting, evidence-based practice, and interprofessional collaboration. Care providers and researchers involved are nurses, occupational therapists, speech therapists, physical therapists, and family doctors.

Steven M. Boker PhD is Professor of Psychology and Director of the Human Dynamics Lab at the University of Virginia, Charlottesville, Virginia, USA. Dr. Boker is team leader of the OpenMx Structural Equation Modeling software project, and developer or co-developer of several methods for fitting differential equations to repeated observations data. He is a Fellow of the Association for Psychological Science and the American Psychological Association and is 2013–14 President of the Society of Multivariate Experimental Psychology.

Ulf G. Bronas PhD is Assistant Professor and Director of the School of Nursing Laboratory of Clinical Physiology at the School of Nursing, University of Minnesota, Minneapolis. His program of research is focused on the cardio-renal protective mechanisms of lifestyle interventions in patients with chronic kidney disease, diabetes, and hypertension. He is recipient of the Jay D. Coffman Young Investigator Award, Fellow of the Society for Vascular Medicine, and Fellow of the American Heart Association.

Gregory G. Brown PhD is Professor of Psychiatry and Director of the NeuroImaging and Behavioral Analysis Laboratory at the University of California, San Diego. He is Editor-in-Chief of *Neuropsychology* and a Fellow of the American Psychological Association. His research involves translational brain imaging of neuropsychological disorders and associated mouse models. His human brain imaging research focuses on the integration of computational models of cognition with functional brain imaging data.

Angela Bryan PhD is Professor of Psychology and Neuroscience, University of Colorado Boulder, Colorado, USA. She is an associate editor for *Annals of Behavioral Medicine* and teaches courses in research methods and statistics. Her research takes a transdisciplinary approach to understanding health behavior, incorporating theoretical models of behavior, neurocognition, genetics, and epigenetics.

Krzysztof J. Cios is Professor and Chair of Computer Science, Virginia Commonwealth University, Richmond, USA, and is affiliated with the Institute of Theoretical and Applied Informatics of the Polish Academy of Sciences. His research interests are in data mining (co-authored the first monograph on data mining on the US market), machine learning, and biomedical informatics. Dr. Cios has been the recipient of the Norbert Wiener Outstanding Paper Award, the Neurocomputing Best Paper Award, and the Fulbright Senior Scholar Award. He is a Foreign Member of the Polish Academy of Arts and Sciences.

Sean P. Clarke RN, PhD, FAAN is Associate Dean for Undergraduate Programs and Professor, William F. Connell School of Nursing at Boston College. His research foci are hospital safety, nursing management, and health policy.

Donna L. Coffman PhD is Research Associate Professor of Health and Human Development and Senior Research Associate at the Methodology Center at Pennsylvania State University. Her research focuses on causal inference for observational studies and mediation analysis.

Yvette P. Conley PhD is Professor of Nursing and Human Genetics at the University of Pittsburgh. Her research focuses on the contribution of genomic and epigenomic variability to patient outcomes.

Sylvie Cossette RN, PhD is Professor at the Faculty of Nursing, University of Montreal and Researcher at the Montreal Heart Institute Research Center, Montreal, Quebec, Canada. She is Co-director of the Quebec Nursing Intervention Research Network.

Nicky Cullum DBE, PhD, RN, FAAN, FMedSci is Professor of Nursing at the University of Manchester and Co-ordinating Editor of the Cochrane Collaboration Wounds Group. Her areas of expertise are wound care, randomized controlled trials, systematic reviews, and meta-analysis.

Henrica de Vet PhD is Professor in Clinimetrics, Department of Epidemiology and Biostatistics, EMGO+ Institute for Health and Care Research, VU University Medical Center, Amsterdam, the Netherlands. Her research is focused on the quality of measurements with applications in the field of musculoskeletal disorders, quality of life in the elderly and end of life care, and PET scans for non-Hodgkin's lymphoma.

Pascal R. Deboeck PhD is Assistant Professor at the University of Kansas. His area of expertise is the development and application of methods for the analysis of intraindividual time series. In particular, he focuses on advancing the use of derivatives, differential equation modeling, and dynamical systems concepts in the social sciences.

Connie White Delaney PhD, RN, FAAN, FACMI is Professor and Dean, School of Nursing, University of Minnesota. She serves as Associate Director of the CTSI-BMI in the Academic Health Center. Dr. Delaney is the first Fellow in the College of Medical Informatics to serve

as a Dean of Nursing. She is an appointee to the US Health Information Technology Policy Committee. She is an active researcher and writer in the areas of national standards development for essential nursing care and outcomes/safety data. She holds a BSN with majors in nursing and mathematics, MA in Nursing – Adult Health, PhD in Educational Administration and Computer Applications, and completed postdoctoral study in nursing and medical informatics at the University of Utah.

Molly C. Dougherty PhD, RN is Professor Emerita, School of Nursing at the University of North Carolina Chapel Hill. She was Editor of *Nursing Research* 1997–2012. Her interests include women's health and globalization of the nursing literature.

Louise Doyle RPN, BNS, MSc, PhD is Assistant Professor of Mental Health Nursing, School of Nursing and Midwifery, Trinity College, Dublin. Her research interests are self-harm and suicidal behavior. She has a particular interest in utilization of mixed methods designs.

Jo Dumville MSc, PhD is Senior Lecturer in Applied Health Research at the University of Manchester. She is an Editor with the Cochrane Wounds Group and has been involved in the conduct of randomized controlled trials, systematic reviews, and meta-analyses.

Patricia Eckardt PhD, MBA, RN is Assistant Professor at Stony Brook University. She is a reviewer for four journals in diverse disciplines. Her research focus is on causal inference with observational studies.

Nancy Feeley RN, PhD is Associate Professor, Ingram School of Nursing, McGill University, Senior Researcher, Centre for Nursing Research and Lady Davis Institute, Jewish General Hospital, Montreal. She is also Co-director of the Quebec Nursing Intervention Research Network.

Julie Fleury PhD FAAN Hanner Professor, College of Nursing and Health Innovation, Arizona State University. Her research is focused on motivation in the initiation and maintenance of health behavior, particularly among vulnerable populations.

Gary H. Glover is Professor of Radiology and director of the Radiological Sciences Laboratory in the Department of Radiology at Stanford University. He has been a pioneer in biomedical imaging technology for nearly 40 years, with contributions in ultrasound, computed tomography, and magnetic resonance imaging, chronicled in some 360 publications. He has received a number of honors, including membership in the National Academy of Engineering, Gold Medal of the International Society for Magnetic Resonance in Medicine (ISMRM), Fellow of the ISMRM, Fellow of the American Institute for Medical and Biological Engineering (AIMBE), and the Outstanding Researcher Award of the Radiological Society of North America (RSNA).

Patricia A. Grady PhD, RN, FAAN is the Director of the National Institute of Nursing Research at the National Institutes of Health. An internationally recognized researcher, her scientific focus has primarily been in the study of strokes. She publishes and lectures on strategic directions, policy, and emerging scientific opportunities in nursing science.

Douglas N. Greve PhD studied computational neuroscience at Boston University prior to joining the Martinos Center at Massachusetts Hospital and the Radiology Department at

Harvard Medical School in 1999. His research on brain image analysis algorithm development has primarily focused on the analysis of fMRI, multimodal integration, and cortical surface-based analysis.

Ulrike Grittner PhD studied mathematics, arts, and educational science at the Humboldt-University in Berlin (Germany). She has worked as a Biostatistician at the Department of Biometrics and Clinical Epidemiology since 2002 and at the Center for Stroke Research at the Charité-University Medical Center, Berlin, Germany since 2012.

Tondi M. Harrison PhD, RN, CPNP is a Principal Investigator in the Center for Cardiovascular & Pulmonary Research at Nationwide Children's Hospital and Assistant Professor, Department of Pediatrics, the Ohio State University, Columbus, Ohio, USA. Her research focuses on maternal-infant interaction and autonomic nervous system development in infants with congenital heart defects.

Nalini Jairath RN, MScN, PhD is Associate Professor, School of Nursing, the Catholic University of America, Washington, DC. She advocates for a more inclusive community of nurse scientists through global outreach and greater openness to those with transdisciplinary education and perspectives.

Nils Lahmann PD Dr MSc, BA, RN is Research Associate at the Charité – University Medical Center Berlin, Germany. He became a trustee of the European Pressure Ulcer Advisory Panel in 2008 and is a member of the EPUAP scientific committee.

Lei Liu PhD is Associate Professor of Biostatistics in the Department of Preventive Medicine and Robert H. Lurie Comprehensive Cancer Center, Northwestern University. He has broad research interests in longitudinal data analysis, survival analysis, and smoothing regression methods.

Thomas T. Liu PhD is Professor of Radiology and Bioengineering and Director of the Center for Functional Magnetic Resonance Imaging (fMRI) at the University of California San Diego. His research is in the area of functional neuroimaging, with a focus on the development of acquisition and analysis methods for fMRI.

Yue Liu MSc is a PhD student of Dr. Lei Liu, Department of Statistics, University of Virginia. Her research interest is the joint modeling of longitudinal and survival processes, especially from a latent class perspective. She has a particular interest in application of latent class joint modeling in clinical studies.

Jessica M. McIlvane PhD is currently Health Science Policy Analyst at the National Institute of Nursing Research. Her prior work as Associate Professor at the University of South Florida focused on aging and health, coping with chronic illness and stress, and health disparities.

Karen A. Monsen PhD, RN, FAAN is Associate Professor, University of Minnesota, School of Nursing, Minneapolis, USA. Her research using large data sets focuses on healthcare intervention effectiveness. She is Co-director of the Center for Nursing Informatics and Director of the Omaha System Partnership.

Bryon A. Mueller PhD is Research Assistant Professor and Magnetic Resonance Imaging Physicist in the Department of Psychiatry at the University of Minnesota. His research focus is the application of novel MR imaging methods to the study of the brain.

Ramona Nelson PhD, RN-BC, ANEF, FAAN is Professor Emerita, Slippery Rock University, Slippery Rock, PA, USA. She is a fellow in NLN Academy of Nursing Education. Her latest books include the 5th edition of *Introduction to Computers for Health Professions* (2010) and *Social Media for Nurses* (2013).

Dat T. Nguyen PhD received BS (1991) and MS (1997) degrees in computer science from the School of Science, Vietnam National University, Ho Chi Minh City (HCMC), Vietnam. In 2009 he joined the Department of Computer Science at Virginia Commonwealth University, USA, as a doctoral student and received his PhD in computer science in 2013. His research interests include data mining, machine learning, high-dimension data visualization and algorithms. He now works as Assistant Professor in the School of Science and Technology, Hoa-Sen University, HCMC, and as Visiting Professor at the University of Finance-Marketing, HCMC.

David Rindskopf PhD is Distinguished Professor at the City University of New York Graduate Center, where he holds a joint appointment in the Educational Psychology and Psychology PhD programs. He is a Fellow of the American Statistical Association, and is a former President of the NY Chapter of the ASA, and past Editor of the *Journal of Educational and Behavioral Statistics* (2008–2010).

Donald B. Rubin PhD is John L. Loeb Professor of Statistics at Harvard University where he has been for the last three decades. He chaired the department for nearly 15 years during that time and is especially known for his contributions to missing data and causal inference.

Sarah J. Schmiede PhD is Assistant Research Professor in the Department of Biostatistics and Informatics at the University of Colorado Denver. She serves as the Biostatistician for the College of Nursing. Her research interests include health psychology and latent variable modeling.

Maria Schubert PhD, RN is Clinical Nurse Scientist and head of team patient safety at the Center of Clinical Nursing Science, University Hospital Zurich, Switzerland. Her expertise and area of research focus on links between system factors such as prioritization, rationing of nursing care, work environment, and patient safety outcomes in acute care.

Souraya Sidani PhD, RN is Professor and Canada Research Chair at the School of Nursing, Ryerson University. Her areas of expertise are in quantitative research methods, intervention design and evaluation, and measurement. Her research focuses on evaluating interventions, examining patient preferences for treatments, and refining research methods and measures for determining the clinical effectiveness of interventions.

Klaas Sijtsma PhD is Full Professor of Methods of Psychological Research, Tilburg University, Tilburg, the Netherlands. He specializes in psychometrics, in particular classical test theory, item response theory, outliers and missing data problems in tests, and test and questionnaire construction.

Nancy Staggers PhD, RN, FAAN is Professor, School of Nursing, University of Maryland, Baltimore. As an AMIA NI-WG-recognized nursing informatics pioneer, she has extensive experience in clinical informatics including IT executive positions for large electronic health record projects. Her research program centers on health IT product usability.

Sarah A. Stoddard PhD, RN, CNP is Assistant Professor in Systems Leadership and Effectiveness Science in the School of Nursing, the University of Michigan. Her research examines how social and environmental factors influence the health trajectories of at-risk youth.

Patricia W. Stone PhD, RN, FAAN is the Centennial Professor in Health Policy at Columbia University School of Nursing, Director of the Center for Health Policy and Co-director of the PhD Program. She has led several federally funded projects investigating cost and quality outcomes around patient safety.

Diane Treat-Jacobson PhD, RN, FAAN is Professor at the University of Minnesota School of Nursing and Chair of the Adult and Gerontological Health Cooperative. Her program of research is focused on exercise interventions to improve quality of life and functional outcomes in patients with peripheral artery disease.

Mayuko Uchida-Nakakoji PhD, RN is a consultant in the healthcare industry. She graduated from Columbia University School of Nursing in 2014. She is passionate about improving quality of care for the geriatric population through patient-centered outcomes and cost-effectiveness research.

Howard Wainer is Distinguished Research Scientist at the National Board of Medical Examiners. He is a fellow of the American Statistical Association and American Educational Research Association. Dr. Wainer has been a recipient of many honors, including the Samuel J. Messick Award for Distinguished Scientific Contributions Award from Division 5 of the American Psychological Association, Career Achievement Awards from the National Council on Measurement in Education and from the Psychometric Society. He is interested in data science in all of its aspects, but especially data visualization. He is interested in the use of evidence in all phases of decision-making. His most recent book, his twentieth, is *Medical Illuminations: Using Evidence, Visualization & Statistical Thinking to Improve Healthcare* (Oxford, 2014).

Bonnie L. Westra PhD, RN, FAAN, FACMI is Associate Professor, University of Minnesota, School of Nursing and core faculty in the Institute for Health Informatics, Minneapolis, USA. Her research focuses on use of electronic health record data for predictive modeling of outcomes. She is Director of the University of Minnesota Center for Nursing Informatics.

Elizabeth R. Zell MStat is President and Principal Statistician at Stat-Epi Associates, Inc., Ponte Vedra Beach, Florida, USA. She recently retired from the Centers for Disease Control and Prevention after an illustrious 32-year career focusing primarily on vaccine preventable diseases.

ACKNOWLEDGMENTS

I benefitted from much encouragement and assistance throughout the process of editing the *Handbook* and it is with much appreciation that I acknowledge this support. It has been an honor to work with distinguished scholars from around the world to create a resource for quantitative methods suitable for emerging, priority problems in nursing science, and it is with heartfelt gratitude that I express my appreciation to them. The topical outline for the book arose from the syllabi for my nursing inquiry/methodology seminars at the University of Minnesota, School of Nursing; I'm grateful to my PhD students for their willingness to tackle challenging content and provide feedback about approaches that worked best to mesh complex substantive and methodological concerns. My colleagues Jean Wyman at the University of Minnesota, Barbara Holtzclaw at the University of Oklahoma, and Donna McCarthy at Marquette University cheered the effort as we collaborated on other projects during the course of preparing the book. The superb team at Routledge – Senior Editor Grace McGinnis, Editorial Assistants James Watson and Louisa Vahtrick, and the production staff – has the patience of saints and their commitment to the *Handbook* ensured that it got completed. During my PhD program in psychometric methods, Bob Cudeck (my advisor), Michael Browne, and Stephen DuToit were instrumental in guiding my development as a methodologist, and I continue to benefit from their perspectives. A pre-doctoral fellowship for my PhD program from the National Institute of Nursing Research at the US National Institutes of Health early in my scientific career and subsequent sabbaticals from the University of North Dakota and the University of Minnesota supported study in psychometrics, community health sciences/epidemiology, and informatics/data analytics that served as the foundation for editing the *Handbook*. My husband George Henly was supportive beyond measure.

This page intentionally left blank

FOREWORD

Nursing science is entering a new era influenced by the type and complexity of research questions on human health, illness, and health care delivery. Expertise in quantitative methodologies is essential for framing research questions and hypotheses, appropriate designs, and analytical methods that focus on topics of interest to nursing scientists ranging from measurement of symptoms, behavior, and patient-reported outcomes using new technologies and tools; basic and clinical research on the mechanisms that underlie health, behavior and behavior change, and treatment responses; effectiveness of different interventions to identify best practices at improving health with the least harm; use of approaches that assess health systems, quality of care, and cost-effectiveness of interventions; and translation of interventions for widespread implementation and adoption. In addition, emerging areas of science in the omics and microbiome, biophysical measurement and imaging, informatics technology, and big data are offering new opportunities for nursing scientists and require the use of cutting-edge analytical methods. Today's nursing scientists require knowledge and skill in a broad range of quantitative methodologies that involve data mining and visualization techniques, complex mathematical modeling for patterns in biopsychosocial health and longitudinal processes, and innovative computational methods in order to address critical questions in the science of health.

This first edition of the *Routledge International Handbook for Advanced Quantitative Methods in Nursing Science*, edited by Susan J. Henly, PhD, RN, FAAN, is a transformative contribution to advancing global nursing science at this point in its evolution. It offers the first in-depth resource of quantitative methods for nursing research that integrates theory, design, and statistical methods. It includes both traditional and cutting-edge approaches addressing measurement; multilevel and longitudinal modeling techniques; research designs for observational and randomized trials; e-science methods incorporating big data techniques, omics, and imaging approaches; comparative effectiveness research methods; and economic evaluations.

Susan Henly is ideally qualified to serve as editor of this important handbook. She brings her scientific training as a psychometrician with expertise in measurement problems in nursing research and longitudinal methods, and experience as Methods Director for the innovative Center for Health Trajectory Research at the University of Minnesota School of Nursing. As the former long-term Associate Editor overseeing statistical reviews of manuscripts in *Nursing Research* and now its Editor, she brings a wealth of knowledge regarding quantitative nursing research. The chapters in the *Routledge International Handbook for Advanced Quantitative Methods*

Foreword

in Nursing Research are co-authored by renowned scientists and methodologists from nursing and related disciplines who have unique capability in the topics of their respective chapters. Her deliberate selection of international scientists to author chapters enriches this text and broadens the perspectives related to quantitative methodologies. The topics discussed are widely applicable to addressing research issues within nursing and advancing global nursing science.

From my own work in teaching graduate students who are at the forefront of addressing quantitative questions in the domain of nursing, I have been challenged to find a comprehensive text that goes beyond the basics of nursing research design and incorporates contemporary and advanced methods. The *Routledge International Handbook for Advanced Quantitative Methods in Nursing Research* will serve as the authoritative reference for advanced PhD students, active nursing scientists, journal editors, and policy makers on advanced quantitative research designs and methods that includes classical and cutting-edge approaches.

Considering the changes in nursing research over the past two decades, the *Routledge International Handbook of Advanced Quantitative Methods in Nursing Research* is a timely and significant contribution to global nursing science. The legacy of this handbook will be to shape nursing research and its discoveries leading to new knowledge and advances in the science of health.

Jean F. Wyman, PhD, RN, FGSA, FAAN
Professor and Cora Meidl Siehl Chair in Nursing Research
School of Nursing
University of Minnesota
Minneapolis, MN, USA
Wyman002@umn.edu; 612-624-2132

PREFACE

Nursing science is a complex enterprise. Describing, predicting, explaining, and controlling health processes and outcomes focuses on persons as individuals or as groups; is inherently multivariate and multiscale in scope; involves idiographic (person-specific) and nomological (general) law-finding; and has roots in the life sciences, physical sciences, and behavioral sciences. Following a long developmental phase in the 20th century, nursing science is now a maturing science – and in need of an authoritative source for methodological foundations of nursing research.

Methodology for nursing science was rooted in the research traditions of psychometrics and the social sciences and arose secondarily from epidemiological and biostatistical perspectives. Recently, mixed methods research appeared as a rapprochement to the long and overly quarrelsome debate about the relative usefulness of qualitative and quantitative approaches in nursing research. Information science analytics and data mining are fast becoming new ways to generate knowledge in nursing research.

The contents of the *Handbook* are wide-ranging. Past methodological approaches that have served basic and clinical/translational nursing scientists well are included along with emerging approaches that will shape its future. Methods are described in integration with research questions relevant to the domain of nursing science, from the cellular and molecular levels of biology to wide social systems. Because health and illness are complex phenomena, multivariate methods are emphasized. Technical presentation of models and methods necessarily involves use of mathematics, but at the same time, contributors have created text that provides an engaging and accessible reading experience. Selection of theory-based methods respects the longstanding preference in nursing for theory-based research and fits comfortably with advances in methodology that view quantitative methods as a way to link theory, design, and statistical model when posing, exploring, and answering research questions.

The *Routledge Handbook of Advanced Quantitative Methods in Nursing Research* is intended for a global readership of active nursing researchers engaged in advancing nursing science; PhD students in nursing science; scientific staff of agencies that fund nursing research, including the peer reviewers of proposals submitted to sponsors; and editors and peer reviewers of journals disseminating nursing research results. It also provides essential methodological background for policy makers and clinicians evaluating research results for application in practice.

The level of presentation assumes a firm grasp of basic research methods and classical measurement methods. A working knowledge of basic descriptive and inferential statistics equivalent to

Preface

two semesters of graduate level study is needed. Mastery of mathematics through college algebra is essential; completion of the first course in calculus and an introduction to linear algebra are helpful. The book should be suitable for PhD-level courses that integrate nursing inquiry with study of multivariate observational design; multivariate quasi-experimental and experimental design; and methods for analyzing big data. The *Handbook* may also be of interest to advanced graduate students and researchers in health psychology and allied health disciplines.

SJH

PART I

The domain of nursing science

This page intentionally left blank

1

THE DOMAIN OF NURSING SCIENCE

Patricia A. Grady and Jessica M. McIlvane

Nurses have an increasingly central role in the healthcare system. Changing demographics, an aging population, cultural and racial diversity, increased health disparities, and the explosion of technology are all impacting health and healthcare. In the United States, passage of the Patient Protection and Affordable Care Act (ACA) has increased focus on prevention. These forces synergize to underscore key roles of the nurse. Nurses constitute the largest component of the healthcare workforce, responsible for the healthcare of individuals, families, and communities around the globe. They interact with patients and families around health and healthcare, and help them to navigate often confusing healthcare systems. As scientists, nurses conduct research to improve health. At a time of increasingly complex societal, economic, and health issues, nurses and nurse scientists are well-positioned to identify pressing research questions, lead the way to discovery of innovative solutions, and translate scientific findings to improve health, healthcare, and quality of life.

This chapter discusses the domain of nursing science; elucidates links between nursing science, practice, and policy; and considers future challenges and opportunities. While not meant to be exhaustive, we consider many areas that comprise the domain of nursing science, as well as major initiatives and societal trends shaping its current and future state. The domain of nursing science is far-reaching, addressing the needs and improving the health of diverse individuals, families, and communities. The foundational motivation for nursing science is to establish the scientific basis for clinical practice through basic, clinical, and translational research, with the ultimate goal of improving health and healthcare.

Areas of focus

From bench to bedside, from clinic to community, nurse scientists conduct research addressing a broad range of issues with real-world health implications. Nurse scientists are leaders in biobehavioral research, focusing on symptom management. They investigate multiple health determinants including genomic, physiologic, psychological, familial, cultural, and environmental factors. They use a variety of methods and approaches including basic lab research; clinical intervention studies; translational, implementation, and comparative effectiveness research; and analyses of cost, outcomes, and quality of care.

Health promotion, disease prevention, and improving quality of life in health and illness are the cornerstones of nursing science. Nurse scientists conduct research to promote positive outcomes across the lifespan. They design coping-skills interventions for parents with premature infants; community-based interventions to reduce sexually transmitted infections in adolescents; psychosocial interventions for informal caregivers; and interventions to help families with decision-making in palliative and end-of-life care. They focus on improving symptom management in clinical settings and optimizing self-management of symptoms. They emphasize the active role of patients and families in their own healthcare.

Along with an emphasis on developing and testing behavioral interventions to promote health, nursing science maintains a strong foundation in basic research. Nurse scientists seek to expand knowledge of underlying biological systems, including genetic contributions to symptoms, and to health conditions such as obesity and sleep disorders. Nursing science emphasizes basic behavioral and social science research in determining the predictors of illness and identifying interventions for at-risk groups and the population at large (Rose, 1992).

Nurse scientists play an integral role in transforming the healthcare system. They are on the leading edge of research into patient-centered care, personalized medicine, and genomics, areas of research that will allow treatments and interventions to be tailored to individuals. Today's healthcare system must cope with an increase in multiple chronic conditions and long-term care needs, requiring a greater emphasis on patient care coordination and transitional care to support individuals from hospital to home. Nurse scientists are on the forefront of developing, testing, and implementing new models of care. With an increasing emphasis on providing quality care while reducing unnecessary tests, treatments, and costs, nurse scientists are working to identify the most effective clinical treatments and behavioral interventions through comparative effectiveness research.

Communicating science to key stakeholders is critical. Nurse scientists translate, disseminate, and implement research findings into evidence-based practice; communicate science to practicing nurses, other healthcare practitioners, patients, communities, and policy makers; and inform health policy. Table 1.1 summarizes the areas of focus, populations studied, and methods used in the domain of nursing science.

Guiding initiatives

Many factors shape the state of nursing science. Research priorities set forth by the United States National Institute of Nursing Research (NINR), along with other national and global initiatives, are some of the driving forces behind nursing. These initiatives provide guidance in the current and emerging content, scope, and domain of nursing science, as well as strategies for transforming healthcare. NINR's strategic plan (2011) reflects priorities developed in collaboration with scientists across the globe; its key focal point is "the science of health." To advance the science of health, NINR focuses on: (a) enhancing health promotion and disease prevention; (b) improving quality of life by managing symptoms of acute and chronic illness; (c) improving palliative and end-of-life care; (d) enhancing innovation in science and practice; and (e) developing the next generation of nurse scientists. NINR supports research to improve health across the lifespan, with an emphasis on underserved populations. NINR is committed to promoting health equity and eliminating health disparities by emphasizing culturally appropriate interventions that incorporate views of diverse communities.

In 2010, Sigma Theta Tau (the Honor Society of Nursing) created a global initiative in commemoration of the centennial of Florence Nightingale's death (Beck, Dossey, & Rushton, 2010). Research priorities for this initiative include: promotion of healthy communities through health

Table 1.1 The domain of nursing science

<i>Areas of Focus</i>	<i>Populations</i>	<i>Methods</i>
• Biobehavioral/Biopsychosocial • Symptom Management (e.g., Pain, Fatigue) • Acute and Chronic Illness • Genetics and Genomics • Infant & Maternal Health • Palliative and End-of-Life Care • Health Promotion & Disease Prevention • Self-Management • Caregiving • Quality of Life • Obesity • Communicable & Non-Communicable Diseases • Health Disparities/Health Inequities • Global Health • Person-Centered/Personalized Care • Decision-Making/Risk Behaviors • Evidence-Based Practice • Behavioral and Social Sciences • Technology and Innovation • Interdisciplinary/Interprofessional/Team Science • Research Methodology/Measurement • Ethics • Policy (Local, State, Federal) • Training Nurse Scientists/Clinicians	• Individuals • Families • Communities • Vulnerable Groups • Diverse Groups (inclusive of sex, race/ethnicity, socioeconomic status, age, urban/rural, sexual orientation) • Global Populations	• Intervention • Translation • Dissemination • Implementation • Randomized Controlled Trials • Comparative Effectiveness Research • Cost Analyses • Community-Based Participatory Research

promotion and disease prevention; implementation of evidence-based practice; targeting needs of vulnerable populations; and research capacity development. The Global Health Initiative (GHI), launched by the Obama administration in 2009, seeks to improve global maternal, newborn, child, and family health (KFF, 2012). The GHI sets specific targets for 2014 related to: maternal health; child health; HIV/AIDS; malaria; TB; nutrition; family planning/reproductive health; neglected tropical diseases; and strengthening the health system. The GHI emphasizes resolving global health disparities and achieving health equity.

Common to these initiatives is a focus on health promotion and disease prevention; health inequities, health disparities, and vulnerable populations; and the importance of evidence-based practice and training nurse scientists. These initiatives will continue to influence nurse scientists as they develop their research programs.

Societal trends

The issues emphasized by any science are shaped, in part, by societal trends, demographic shifts, economic circumstances, and political factors. It is within this context that nurse scientists conduct research and affect change. Increasing racial/ethnic diversity coupled with persistent inequities in health call for development, dissemination, and implementation of culturally appropriate interventions for diverse groups.

Populations across the globe are aging rapidly (CDC, 2003; NIA, 2007). Worldwide, the number of people age 65 and older is projected to increase from 500 million in 2006 to one billion in 2030; by 2030, projected increases of up to 140% are anticipated in developing countries (NIA, 2007). Increased numbers of older adults with multiple, chronic conditions (e.g., heart disease, arthritis) will need care that adds significant costs to healthcare systems (NIA, 2007). Burden on informal caregivers will rise (IOM, 2008), as will burden on developing countries with limited resources for basic public health concerns (e.g., maternal and infant health; CDC, 2003).

Challenges in preventable conditions will influence the provision of care and necessitate new ways to intervene. Obesity threatens to increase healthcare costs and decreases quality of life. In 2008, over 1.4 billion adults worldwide were overweight, of whom 500 million were obese (WHO, 2012a). Worldwide, an estimated 34 million people were infected with HIV in 2010, 68% of whom lived in sub-Saharan Africa (The Global Fund, 2012). Prevention of HIV/AIDS will continue to be a global priority.

In the United States, these changes are occurring against the backdrop of an evolving healthcare system. The ACA is stimulating dramatic shifts in healthcare. Expected changes include more people with health insurance; greater emphasis on prevention and wellness; and implementation of new models of care. New technologies (e.g., genetic testing, electronic health records) may improve health outcomes, but healthcare providers must learn to use them, and unforeseen ethical issues in patient care may arise. Disseminating new technologies in developing countries that lack healthcare, economic resources, and infrastructure for research and program delivery is a serious challenge.

This convergence of worldwide trends presents challenges and opportunities as nursing science moves forward in the 21st century. Nurse scientists will continue to play a vital role in health promotion and disease prevention efforts for diverse populations; address pressing issues in evolving healthcare systems; and develop interventions to improve quality of life for everyone.

Multifaceted nature of nursing science

Nursing science is multifaceted in both its domain and methods of research. First, nursing science integrates research questions across biological, psychological/behavioral, and social realms. Second, nursing science addresses research questions across the spectrum of basic, clinical, and translational research. Great strides have been made in establishing a strong foundation of basic and clinical biobehavioral research and in translating findings to inform practice; the following represent some remaining challenges.

Biopsychosocial approaches

Integration of behavior and biology is one of the most central and influential principles of nursing science (Grady, 2006). Nurse scientists have long examined interactions among biological, psychological, and social aspects of health while also considering the sociocultural contexts in which people live. To better understand, prevent, and address complex health problems, such as the co-occurrence of chronic obesity, diabetes, and heart disease, continued emphasis on multiple perspectives and disciplines will be more likely to result in solutions and better patient outcomes (Grey & Connolly, 2008).

A thorough understanding of the complex dynamic relationships between biopsychosocial variables and health requires interprofessional collaborations and a team science approach. Team science can take various forms of collaboration across disciplines including: multidisciplinary – research coordinated among scientists from different disciplines; interdisciplinary – research by a group of scientists on issues that fall on the boundaries of

various disciplines; and transdisciplinary – researchers from different disciplines working towards a common scientific goal involving information exchange and the integration and modification of scientific approaches (Grady, 2010). Each has benefits and ways of operating that need to be navigated by the research group.

Interdisciplinary collaborations provide a richness of theoretical perspectives, intervention strategies, and application of innovative technologies that would not be possible with one discipline (Bindler, Richardson, Daratha, & Wordell, 2012; Grey & Connolly, 2008). Such collaborations increase understanding of complicated issues and increase chances of changing practice and policy. However, interdisciplinary research can be time consuming, expensive, and labor intensive. It can be difficult to integrate varying terminologies, theoretical perspectives, and research methods across disciplines and may present challenges for scientists to obtain funding and publish in traditional venues. Varying cultural norms, practices, and regulations across institutions may present additional barriers (Grey & Connolly, 2008). Yet, interdisciplinary collaborations provide great potential to improve patient outcomes, and researchers can reap the rewards of such collaborations with careful planning, agreement on shared and independent responsibilities, clear communication, and building trust within the research team (Bennett & Gadlin, 2012; Bindler et al., 2012).

There are many examples of integrated biopsychosocial perspectives in research. Dr. Pamela Mitchell and colleagues designed an intervention that taught people recovering from stroke to recognize depression as observable and modifiable, and to increase the level of pleasant events to improve mood. In an efficacy RCT over the year following a stroke event, they found a significant difference at 12 months in reduction of depressive symptom severity in those receiving the experimental treatment and antidepressants, compared with those receiving usual care and antidepressants (Mitchell et al., 2009). This work on post-stroke depression illustrates links between psychological/behavioral factors and physical illness and the important role clinicians play in facilitating recovery from stroke.

Numerous examples of biobehavioral nursing science arise from the rapidly growing area of genomics. Millions of Americans of all ages suffer from pain, fatigue, and other distressing symptoms. Such symptoms are often not well understood, and few effective treatments are available. NINR intramural researchers are seeking to uncover biological causes, and develop the best diagnostic tools and treatments for these distressing symptoms. Dr. Xiao Min Wang's research found a correlation between pain intensity in oral surgery patients and gene expression of cytokines and chemokines at the site of inflammation (Wang et al., 2009). Dr. Leo Saligan's research group is examining molecular-genetic mechanisms of cancer-related fatigue in patients receiving radiation therapy. They found a relationship between changes in expression of eight genes and increased self-reported, cancer-treatment-related fatigue over time (Hsiao, Wang, Kaushal, & Saligan, 2013). In genomic studies such as these, a better understanding of potential chemical "targets" in the body could enhance understanding of biobehavioral influences on health, lead the way to new drug and behavioral treatments, and potentially identify individuals at risk for certain conditions that might benefit from early intervention.

Translational science

Identifying relationships among biological, psychological, and social factors in basic and clinical research is a necessary first step. Translational research then builds the bridge from "bench to bedside" that transforms findings from research into new clinical tools, processes, or applications (Grady, 2010). Nurse scientists have been on the leading edge in bringing person-centered, translational research to the forefront of the health sciences.

Nursing science has built a strong foundation in basic and clinical research that serves to support the early stages of translational research. However, efforts to translate findings into effective, generalizable interventions that can be implemented across various settings have been less successful (IOM, 2001). For example, much of the work in genomics has been basic laboratory-based research, providing a greater foundation for understanding complex biopsychosocial interactions and health outcomes. A greater emphasis is needed on translating basic science findings to improve health (Calzone & Jenkins, 2011; Conley & Tinkle, 2007).

The clinical expertise of nurses enlightens and enriches translation efforts. Sharon Morris, an operating room nurse, recognized that retained objects in patients following surgery was a continuing hazard. Morris partnered with colleagues to develop a prototype system that tagged surgical sponges with radio frequency identification (RFID) chips that could be detected with a handheld device in real-time in the operating room (Macario, Morris, & Morris, 2006). The handheld wand device proved to accurately and rapidly detect RFID-tagged sponges and resulted in the FDA-approved SmartSponge System[®] manufactured by ClearCount Medical Solutions. The SmartSponge System[®] is helping to improve patient safety and save lives in operating rooms throughout the United States. In this example, Morris saw a critical need in clinical practice, worked with an interdisciplinary team to identify a novel solution, and devised a tool with real-world applications in reducing preventable medical errors.

Another example of translating basic nursing science into clinically relevant applications is found in research on wound healing. Chronic wounds are a serious clinical problem affecting people across the lifespan, including children suffering trauma, veterans with combat injuries, and diabetics with non-healing ulcers. From her prior experiences working as a pediatric nurse in a burn unit, Dr. Marjana Tomic-Canic brought a unique perspective to basic, clinical, and translational research on wound healing. Tomic-Canic and colleagues have explored the development of a technique using genetic markers in the skin that could serve as biomarkers for guiding the treatment of wounds in real-time (Tomic-Canic et al., 2008). This technique, which produces data that look like colored product “bar codes,” could be used to better define wound margins, identify non-healing edges, and serve as clinical tools to guide wound debridement. This basic science research has the potential to improve health outcomes and reduce healthcare costs. This example demonstrates how nurse scientists may identify a need in the clinical setting and design a research study testing novel solutions for real-world problems to improve patient care.

Linking science with practice and policy

Clinical practice and nursing science

One of the hallmarks of nursing science is its potential to change clinical practice to improve patient outcomes. Nursing science informs and influences practice in three ways: (a) at the point of care, (b) at the system level, and (c) through dissemination and implementation. Alone and in interaction, they are key to the ultimate goal of moving basic, clinical, and translational research into the practice arena.

Point of care

Nursing science informs practice at the point of care when nurses use scientific evidence to inform their decisions in treating individual patients or groups of patients. As such, evidence-based practice involves making treatment decisions by combining the best scientific evidence with clinical expertise, while also considering patient wishes and preferences (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996).

Evidence-based practice has potential to improve outcomes and reduce costs, however, implementing it is not without challenges (Melnik, Fineout-Overholt, Gallagher-Ford, & Kaplan, 2012). For instance, how do practicing nurses access and use the most current, valid scientific information? How do they address conflicting research findings? Are there educational or organizational barriers? Scientists should consider these issues early in the research process to facilitate identification of the most relevant information for practicing nurses and the most effective ways to communicate scientific findings.

One challenge is the lack of consensus on what constitutes “valid” information in evidence-based practice. Traditionally, evidence from randomized controlled trials (RCTs) has been considered the gold standard. However, RCTs are not always possible or appropriate to answer a particular research question and other research methods (e.g., quasi-experimental, qualitative) may provide valuable information for evidence-based practice.

A study of over 1,000 nurses from across the United States revealed generally positive perceptions about, and willingness to use, evidence-based practice (Melnik et al., 2012). However, barriers were identified, including lack of time, mentors, and organizational support, as well as the need for more training and a clearinghouse of information and research findings. An additional barrier may be that some educational programs focus on teaching rigorous research methods while neglecting instruction on how to apply research evidence in patient care.

System level

Nursing science informs practice when a new or revised practice (e.g., pain management guidelines) is implemented in a particular institution, organization, or setting (e.g., Magnet hospital). The challenge is to figure out how best to implement the new practice in light of the organizational culture, structure, leadership, and attitudes about the new or revised practice. Introducing and implementing new or revised evidence-based practices at the system level can be difficult. Raising awareness and promoting positive attitudes about the changes need to be addressed before focusing on changing behavior and integrating new behaviors into daily practice (Cullen & Adams, 2012). Such strategies must consider the organizational culture and potential barriers to change. While there is a growing body of knowledge in this area, there are few RCTs that have examined implementation of new practices across various healthcare settings.

Dissemination and implementation

Once there is evidence that an intervention is effective, how does one let healthcare providers know about it (dissemination) and how does one integrate the intervention into standard practice (implementation)? Dissemination research addresses the questions: “how, when, by whom, and under what circumstances research evidence spreads throughout the agencies, organizations, and front line workers providing public health and clinical services” (NIH, 2009). Implementation research is “the scientific study of methods to promote the integration of research findings and evidence-based interventions into healthcare policy and practice. It seeks to understand the behavior of healthcare professionals and support staff, healthcare organizations, healthcare consumers and family members, and policymakers in context as key variables in the sustainable adoption, implementation and uptake of evidence-based interventions” (NIH, 2009).

One important consideration for dissemination is to involve community members and stakeholders in the early stages of research. Community-based participatory research (CBPR) incorporates the ideas of community members at all points in the research process, from early identification of research questions, to study design, to intervention implementation (OBSSR,

2012). Part of the CBPR strategy is to get “buy-in” from community leaders, to involve them in identifying challenges, and to get their input in designing the research/intervention to address those challenges. This helps ensure that interventions are culturally relevant to the communities where they will be implemented.

Implementation often involves striking a balance between demonstrating that “core components” of the intervention were implemented as intended while at the same time allowing for flexibility in “adapting” some components to match the needs of a particular setting (Rabin et al., 2008; Damschroder et al., 2009). In the early stages of research planning, it may be helpful to think about how an intervention can be implemented across diverse settings and what aspects can be altered to fit the needs of a particular setting while still maintaining the integrity of the intervention. Interdisciplinary and transdisciplinary collaborations are needed, including reliance on stakeholders in clinical and community settings, as well as consumers and family members. Key variables that determine whether the intervention will be successfully disseminated and implemented include: characteristics of the intervention (e.g., ease of administration, cost); characteristics of the adopters (e.g., individuals/nurses, organizations, communities); and contextual factors (e.g., political, organizational, leadership factors) (Damschroder et al., 2009). Keeping these factors in mind in the initial stages of intervention design may facilitate the later successful dissemination and implementation of evidence-based interventions in real-world settings.

Policy and nursing science

Just as nursing science and practice synergize, science and policy have a reciprocal relationship. On the one hand, policy decisions can have a profound impact on healthcare and clinical practice (e.g., Medicare, ACA). On the other hand, nurses and nurse scientists have the opportunity to inform or even change health policy by communicating their clinical experiences and scientific findings to policymakers. Nurse scientists are currently doing research that informs and changes health policy at the local level (e.g., health insurance and hospital policies), the state level (e.g., state policies on patient-nurse ratios in hospitals), the community level (e.g., school-based physical activity programs), and the federal level (e.g., ACA).

Several characteristics of nursing research make the information it generates particularly useable by policymakers: nursing research addresses public health issues; focuses on people, families, and communities; and draws upon the richness of perspectives from multiple disciplines through a team science approach (Hinshaw, 2011). To do the research alone will not affect change. To change clinical practice and policy, nurse scientists must also ensure that research findings reach the intended audiences.

On occasion, research will influence policy in a spontaneous way because findings are relevant to an issue that is “hot” or in the national spotlight; such findings will be of interest to policymakers at that moment in time. More often, nurse scientists need to plan long-term for how their research might inform policy. Identifying a gap and using rigorous methods are only the first steps; researchers must keep policy implications in mind when they design their study (Grady, 2011). A research project might be successful and produce important, practical results. However, if the intervention is too complicated, costly, or difficult to implement (e.g., unreasonable limits on lifestyle, side effects), then it is unlikely to be amenable to widespread implementation. Nurse scientists must think ahead about the policy issues that might be around the corner, design studies that will have widespread influence, involve important stakeholders early in the process, and clearly communicate research findings in useable ways to policymakers.

The following research examples demonstrate the integral relationships among research, practice, and policy, and illustrate how successful long-term programs of research can inform practice and policy in settings including hospital, home, and community. As part of a long-term research program addressing nursing workforce issues, Dr. Linda Aiken's research team (2010) reported an association between legislatively mandated nurse staffing, nurse job satisfaction, and patient mortality. Policymakers heard a powerful message: adequate hospital nurse staffing improves quality of care and saves lives. Aiken's research has influenced nurse staffing policies in the United States and abroad. Dr. Mary Naylor's long-term program of research established the effectiveness of nurse-managed transitional care in reducing rehospitalizations, improving patient outcomes, and reducing costs (e.g., Naylor et al., 2004). Policy implications were explicated (Naylor & Kurtzman, 2011) and disseminated to clinicians, insurers, and policy leaders; now, the ACA contains provisions for transitional and coordinated services. Dr. Loretta Sweet Jemmott's intervention research documented that culturally appropriate interventions reduce risky behaviors associated with HIV/AIDS and other infectious diseases in African American and Latino adolescents (e.g., Jemmott, Jemmott, Braverman, & Fong, 2005; Jemmott, Jemmott, Fong, & Morales, 2010). Her interventions were subsequently identified as "best-evidence interventions" by the U.S. Centers for Disease Control; strong collaborative community engagement throughout the research process contributed to implementation and acceptance by communities, neighborhoods, and families.

Nursing science and global health

Globalization presents challenges, and also new opportunities to form international research collaborations with the potential to strengthen the worldwide impact of scientific discoveries for improved health. Nurse scientists can lead the way in developing international research collaborations; creating the evidence base for practice; developing low cost, easily administered interventions with broad application; and disseminating and implementing interventions across diverse communities.

Developed nations with varied health systems and values face common issues such as needing to reduce chronic illness burden, providing quality healthcare at lower costs, eliminating health disparities, and integrating new technologies. Nurse scientists will need to adapt and adjust research questions and solutions based on the context and pressing issues of each nation; international collaboration requires an awareness and respect for the uniqueness of how science is done in different settings.

Developing countries face unique and often urgent public health issues. While developed countries experienced an epidemiological shift from infectious to chronic disease as the main threats to health, poor, developing countries are seeing a "triple burden" of communicable disease (e.g., TB, HIV/AIDS), non-communicable disease (e.g., cardiovascular disease), and sociobehavioral illness (e.g., depression) (WHO, 2012b). This poses new challenges for nurse scientists and requires an examination of new approaches to intervention. Again, scientific endeavors could benefit from a CBPR approach involving community stakeholders early in the research process, integrating the wisdom of local communities, and eliciting local input at every step of the research process (OBSSR, 2012).

As an example, Dr. Samuel Sia and colleagues (Chin et al., 2011), supported in part by the NINR, developed and tested a portable, point-of-care, "lab-on-a-chip" diagnostic device to detect HIV and associated co-infections. This battery-powered, disposable, highly accurate, credit card-sized diagnostic device can produce blood-based diagnoses within minutes in practically any healthcare setting; it could be invaluable to HIV prevention efforts in developing countries, or in underserved areas across the world.

Future of nursing science

Nurse scientists are poised to lead the way in transforming the healthcare system; working in and leading interdisciplinary teams; translating, disseminating, and implementing science to inform practice and policy; and ultimately improving patient outcomes. As we look to the future, it behooves us to reflect on priorities for research and training, as well as ways nurse scientists can be prepared as savvy researchers in rapidly changing healthcare and global landscapes.

Multifaceted nursing science

Nurse scientists will face challenges and opportunities as they continue to be leaders in biobehavioral research. Here are some suggestions for how to prepare for the research challenges of the future:

The complexity of issues studied by nurse scientists calls for rigorous research training and a comprehensive, varied toolbox of multivariate statistical and research methods. This handbook is a good starting point in building the toolbox as it provides in-depth information on quantitative methods to prepare nurse scientists for the future. Pre- and post-doctoral programs will need to prepare nurse scientists with the skills needed to address future health problems.

Nurse scientists should seek training opportunities to learn new methods and technologies. Examples are NINR's Summer Genetics Institute and their methodology boot camp focusing on symptom science.

Nurse scientists at all levels should continue to build interdisciplinary and transdisciplinary research teams to address multifaceted and complex health problems (Grey & Connolly, 2008).

Academic and research cultures should change to better support junior scientists taking on interdisciplinary research projects that are time consuming, expensive, and require shared resources. Such efforts should be valued in tenure decisions, publishing, and grant funding.

Informing practice and policy

One of the biggest challenges for nursing and health research is to transition from merely collecting data to positively changing clinical practice, health policy, quality of care, and ultimately health outcomes. Areas to consider as we move the field forward are:

Junior scientists need to think several steps ahead when they develop a program of research. Strategies to translate, disseminate, and implement interventions and inform evidence-based practice, as well as anticipation of barriers and opportunities, need to be considered early in the research process. They may not be personally engaged in each of these steps, but their planning will be important in facilitating the steps.

The above suggestion also applies to policy. Nurse scientists need to think through policy implications of their work, and even develop research specifically to inform policy. It will be beneficial to invest in relationships with policymakers at all levels (e.g., hospital, state, federal) and communicate scientific findings in clear, meaningful ways.

Publishing research findings in scientific journals is necessary but not sufficient. Long-term programs of research should also aim for translating, disseminating, and implementing findings to inform practice and policy and optimize patient care.

As a field, nursing science must invest in today's students and tomorrow's leaders in science and policy and open new doors for partnerships at the local, state, federal, and international levels (Grady, 2011). Understanding the different ways in which science informs practice (e.g., at point of care and system levels) should be a key aspect of training.

Nurse scientists should seek out additional resources available to scientists in general (e.g., Baron, 2010) and nurse scientists in particular (e.g., Hinshaw & Grady, 2011) on how to communicate science in meaningful ways to policymakers.

The *International Handbook of Advanced Quantitative Methods for Nursing Research* is timely in that it provides nurse scientists with valuable research and statistical skills that will serve them well as they study complex issues, develop interventions to promote health and prevent disease, and mentor the next generation of nurse scientists in a rapidly changing world and healthcare landscape. We hope this broad overview of nursing science sets the stage for the eloquent discussions of research and statistical methods in the remainder of the book, and that it provides useful information to junior and senior scientists in their career development and research efforts.

References

- Aiken, L. H., Sloane, D. M., Cimiotti, J. P., Clarke, S. P., Flynn, L., Seago, J. A., . . . Smith, H. L. (2010). Implications of the California nurse staffing mandate for other states. *HSR: Health Services Research*, 45, 904–921.
- Baron, N. (2010). *Escape from the ivory tower: A guide to making your science matter*. Washington, DC: Island Press.
- Beck, D. M., Dossey, B., & Rushton, C. H. (2010). The 2010 international year of the nurse initiative. *American Journal of Nursing*, 110, 11.
- Bennett, L. M., & Gadlin, H. (2012). Collaboration and team science: From theory to practice. *Journal of Investigative Medicine*, 60, 768–775.
- Bindler, R. C., Richardson, B., Daratha, K., & Wordell, D. (2012). Interdisciplinary health science research collaboration: Strengths, challenges, and case example. *Applied Nursing Research*, 25, 95–100.
- Calzone, K. A., & Jenkins, J. (2011). Genomics education in nursing in the United States. *Annual Review of Nursing Research*, 29, 151–172.
- Centers for Disease Control and Prevention (CDC). (2003). Public health and aging: Trends in aging—United States and worldwide. *Morbidity & Mortality Weekly Report*, 52, 101–104, 106.
- Chin, C. D., Laksanasopin, T., Cheung, Y. K., Steinmiller, D., Linder, V., Parsa, H., . . . Sia, S. K. (2011). Microfluidics-based diagnostics of infectious diseases in the developing world. *Nature Medicine*, 17, 1015–1019.
- Conley, Y. P., & Tinkle, M. B. (2007). The future of genomic nursing research. *Journal of Nursing Scholarship*, 39, 17–24.
- Cullen, L., & Adams, S. L. (2012). Planning for implementation of evidence-based practice. *Journal of Nursing Administration*, 42, 222–230.
- Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implementation Science*, 4, 50.
- Grady, P. A. (2006). Biobehavioral research at NINR and NIH. *Nursing Outlook*, 54, 300–302.
- Grady, P. A. (2010). Translational research and nursing science. *Nursing Outlook*, 58, 164–166.
- Grady, P. A. (2011). Research: A foundation for health policy. In A. S. Hinshaw & P. A. Grady (Eds.), *Shaping health policy through nursing research* (pp. 17–33). New York, NY: Springer.
- Grey, M., & Connolly, C. A. (2008). “Coming together, keeping together, working together”: Interdisciplinary to transdisciplinary research and nursing. *Nursing Outlook*, 56, 102–107.
- Hinshaw, A. S. (2011). Science shaping health policy: How is nursing research evident in such policy changes? In A. S. Hinshaw & P. A. Grady (Eds.), *Shaping health policy through nursing research* (pp. 1–15). New York, NY: Springer.
- Hinshaw, A. S., & Grady, P. A. (Eds.). (2011). *Shaping health policy through nursing research*. New York, NY: Springer.
- Hsiao, C. P., Wang, D., Kaushal, A., & Saligan, L. (2013). Mitochondria-related gene expression changes are associated with fatigue in patients with nonmetastatic prostate cancer receiving external beam radiation therapy. *Cancer Nursing*, 36, 189–197.
- Institute of Medicine (IOM). (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academies Press.
- Institute of Medicine (IOM). (2008). *Retooling for an aging America: Building the health care workforce*. Washington, DC: National Academies Press.

- Jemmott, J. B., III, Jemmott, L. S., Braverman, P. K., & Fong, G. T. (2005). HIV/STD risk reduction interventions for African American and Latino adolescent girls at an adolescent medicine clinic: A randomized controlled trial. *Archives of Pediatrics & Adolescent Medicine*, 159, 440–449.
- Jemmott, J. B., III, Jemmott, L. S., Fong, G. T., & Morales, K. H. (2010). Effectiveness of an HIV/STD risk-reduction intervention for adolescents when implemented by community-based organizations: A cluster-randomized controlled trial. *American Journal of Public Health*, 100, 720–726.
- Kaiser Family Foundation (KFF). (2012). Fact sheet: The US global health initiative (GHI). Retrieved from www.kff.org/globalhealth/upload/8116-03.pdf
- Macario, A., Morris, D., & Morris, S. (2006). Initial clinical evaluation of a handheld device for detecting retained surgical gauze sponges using radiofrequency identification technology. *Archives of Surgery*, 141, 659–662.
- Melnyk, B. M., Fineout-Overholt, E., Gallagher-Ford, L., & Kaplan, L. (2012). The state of evidence-based practice in US nurses: Critical implications for nurse leaders and educators. *Journal of Nursing Administration*, 42, 410–417.
- Mitchell, P. M., Veith, R. C., Becker, K. J., Buzaitis, A., Cain, K. C., Fruin, M., . . . Teri, L. (2009). Brief psychosocial behavioral intervention with antidepressant reduces poststroke depression significantly more than usual care with antidepressant. Living Well with Stroke: Randomized, controlled trial. *Stroke*, 40, 3073–3078.
- National Institute of Nursing Research (NINR). (2011). *Bringing science to life: NINR strategic plan*. Retrieved from www.ninr.nih.gov/AboutNINR/NINRMissionandStrategicPlan/
- National Institute on Aging (NIA). (2007). *Why population aging matters: A global perspective*. Retrieved from www.nia.nih.gov/sites/default/files/WPAM.pdf
- National Institutes of Health (NIH). (2009). Dissemination and implementation research in health (R01), program announcement 10–038. Retrieved September 26, 2012, from <http://grants.nih.gov/grants/guide/pa-files/PAR-10-038.html>
- Naylor, M. D., Brooten, D. A., Campbell, R. L., Maislin, G., McCauley, K. M., & Schwartz, J. S. (2004). Transitional care of older adults hospitalized with heart failure: A randomized, controlled trial. *Journal of the American Geriatrics Society*, 52, 675–684.
- Naylor, M. D., & Kurtzman, E. T. (2011). Transitional care: Improving health outcomes and decreasing costs for at-risk chronically ill older adults. In A. S. Hinshaw & P. A. Grady (Eds.), *Shaping health policy through nursing research* (pp. 201–213). New York, NY: Springer.
- Office of Behavioral and Social Sciences Research (OBSSR). (2012). Community-based participatory research. Retrieved from http://obssr.od.nih.gov/scientific_areas/methodology/community_based_participatory_research/index.aspx
- Rabin, B. A., Brownson, R. C., Haire-Joshu, D., Kreuter, M. W., & Weaver, N. L. (2008). A glossary for dissemination and implementation research in health. *Journal of Public Health Management and Practice*, 14, 117–123.
- Rose, G. (1992). *The strategy of preventive medicine*. Oxford: Oxford University Press.
- Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal*, 312(7023), 71–72.
- The Global Fund. (2012). Fighting AIDS: The global AIDS epidemic. Retrieved from www.theglobalfund.org/en/about/diseases/hiv/aids/
- Tomic-Canic, M., Ayello, E. A., Stojadinovic, O., Golinko, M. S., & Brem, H. (2008). Using gene transcription patterns (bar coding scans) to guide wound debridement and healing. *Advances in Skin & Wound Care*, 21, 487–492.
- Wang, X. M., Hamza, M., Wu, T. X., & Dionne, R. A. (2009). Upregulation of IL-6, IL-8 and CCL2 gene expression after acute inflammation: Correlation to clinical pain. *Pain*, 142, 275–283.
- World Health Organization (WHO). (2012a). Obesity and overweight. Retrieved from www.who.int/mediacentre/factsheets/fs311/en/index.html
- World Health Organization (WHO). (2012b). Health transition. Retrieved from www.who.int/trade/glossary/story050/en/index.html

2

THEORIZING IN NURSING SCIENCE

Susan J. Henly

Efforts to develop nursing theory proceeded with collective self-awareness amidst vigorous debate from which metatheory (Fawcett, 1984), grand theory (Marriner-Tomey & Alligood, 2002) and middle range theory (Smith & Liehr, 2008) emerged. Foundational papers documenting this largely philosophical endeavor were collected in Nicholl (1986); continued relevance for the discipline and practice of nursing was argued in Chinn and Kramer (2011); and a remapping of the philosophical orientation for contemporary nursing was proposed by Risjord (2010). From these efforts, theory has come to a valued and privileged position in the heritage of nursing as an academic discipline.

The focus of this chapter is theorizing in contemporary nursing science. The concept map in Figure 2.1 depicts theorizing as an iterative process of integrating nursing inquiry with classical and e-science methods used to investigate nursing phenomena; it reflects the three pillars of contemporary science (theory, modeling, and computation; President's Information Technology Advisory Committee, 2005). The perspective is pragmatic, emphasizing theorizing as it is actually done in the day-to-day work of scientists (cf., Blum, 2014; Hoffmann, 2003). The nursing metaparadigm (health, person, environment, nurse, time) is central. Person-centered idiographic theory and general law-focused nomothetic theory are distinguished. Expressing scientific theory as a statistical model that can be estimated and evaluated and knowledge-building using big data-based e-science approaches are addressed. The final stage of nursing research – translation of findings to clinical practices (Grady, 2010) – optimizes processes and outcomes of care and reveals directions for future research.

Nursing science is the science of health

Nursing science is the science of health (Donaldson, 2003; Grady & McIlvane, this volume; NINR, 2011). Nursing science focuses on health and illness in individuals, families, and communities, in ecological context and over time across lifetimes and generations. Design and evaluation of self-care and nursing interventions constitute an area of unique emphasis; intervention goals are aimed at health promotion, disease prevention, symptom mitigation, and provision of compassion at end of life. Nursing intervention research includes articulation of mechanisms that create effects. Further, understanding the impact of variations in nursing systems on quality, safety, economics of care and care systems, and implementation science are integral to nursing research.

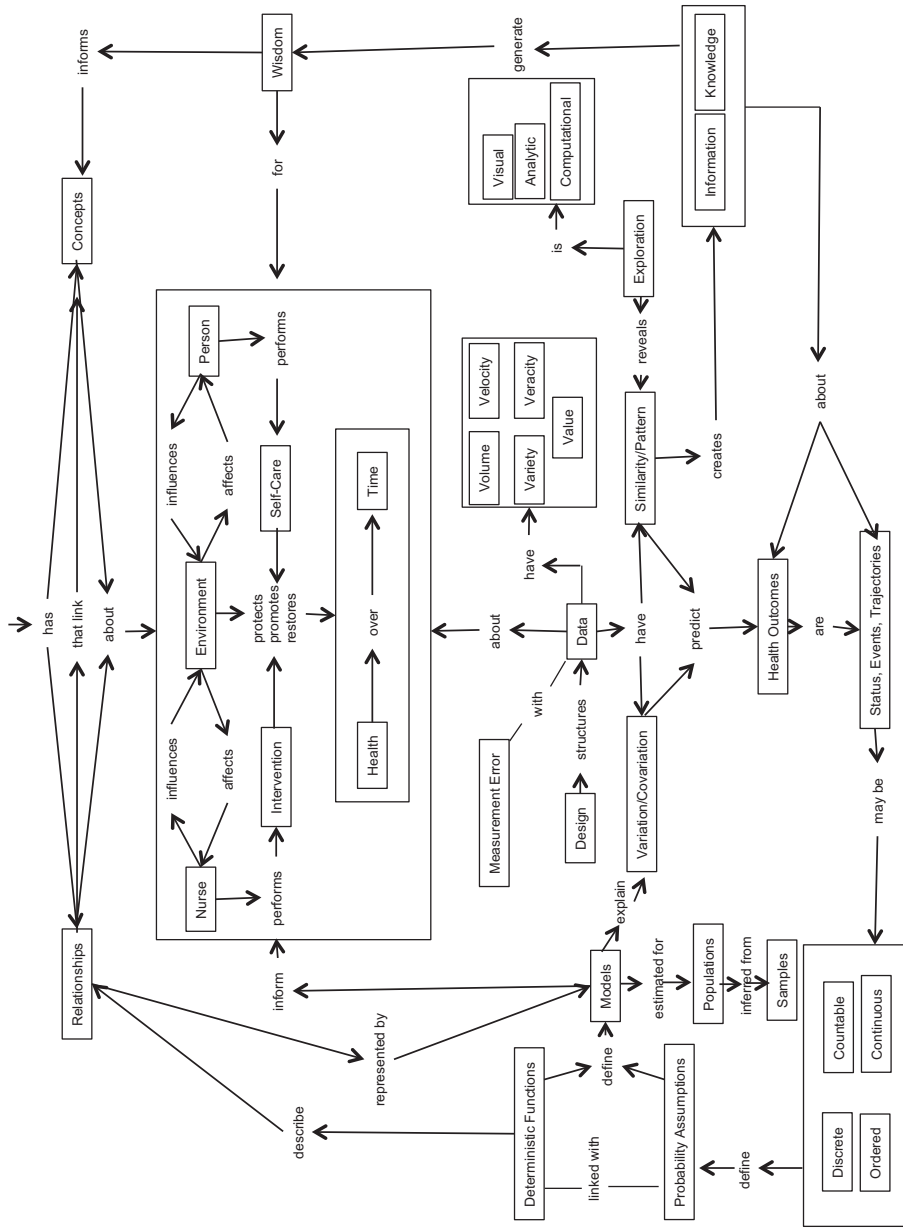


Figure 2.1 Concept map: theorizing in nursing science. The concept map is designed to answer the focus question, “What is scientific nursing theorizing about?” Concepts (labels for perceived regularities in the events and objects of scientific nursing theorizing) are shown in boxes, some of which are nested, and arise from regularities in substantive and methodological domains and events (activities of science), respectively. The concepts within and across substantive and methodological domains are linked by propositions that create semantic units (units of meaning). The concept map demonstrates how the methods of classical science and the knowledge discovery processes of e-science are enmeshed with substantive content and mutually codependent. Concept maps are described in Novak & Cañas (2008).

Nursing metaparadigm and knowledgescape

The metaparadigm of person, environment, health, and nursing (Fawcett, 1984; Gortner, 1983) has been the central organizing structure for nursing theory for over 30 years. Time was added to the metaparadigm (Henly, Wyman, & Findorff, 2011) because it is essential to understanding health and illness, especially the processes by which changes in health status are effected (Donaldson & Crowley, 1978; Gottlieb & Feeley, 1995; Hyman & Corbin, 2001). Inclusion of time allows a person-centered perspective (cf., Molenaar, 2004) in nursing science consistent with practice values (Henly, 2007). As in related disciplines, environment as context is fundamental to understanding the functioning of human beings, individually and in groups, in health and illness across a wide range of spatiotemporal scales (cf., Cairns, Elder, & Costello, 1996; Cohen, 2004; Krieger, 2001; Webster & Ward, 2011).

Concepts

In classical science, concepts are understood as the building blocks of theory (Polit & Hungler, 1999, pp. 24–25, 106–107) arising in concert from substantive phenomena and the methods used to obtain, manipulate, and interpret data. In e-science, the notion of concept is nuanced but can be regarded as a basic unit of thought underlying human intelligence and communication (Yao, 2004) and arising from human knowledge and information at hand (Guergachi, 2003, p. 229).

The person metaconcept (as an individual or group, from molecules to populations) is composed of integrated biological, psychological, biobehavioral, social, and spiritual characteristics (Shaver, 1985) that bear on health and illness at any one point in time and over time and influence differences among persons in health status, health events experienced, and health trajectory (Henly, Wyman, & Findorff, 2011). Persons are viewed holistically, but without disallowing the interacting scales of systems, from molecular to experiential, of which they are composed (i.e., acknowledging the reality of multiple scales is not equivalent to being reductionistic, but is realistic; Founds, 2009; also see Dada & Mendes, 2011). Environment as context for health includes social, biological, and physical qualities experienced in multiscale built and naturally occurring settings; the totality of exposures arising from the environment is now referred to as the *exposome* (Louis & Sundaram, 2012; Rapa-port, 2012). From the time of Nightingale (1860/1969), environmental manipulation has been recognized as a fundamental intervention focus in nursing. Self-care and nursing interventions are actions taken to improve or promote health or prevent disease and constitute the special class of variables that are deliberately manipulated to reach a health-related goal (Melnik & Morrison-Beedy, 2012; Sidani & Fleury, this volume). Time and the measurement of time reflect clock/calendar, biological/social, perceived time and transcendence dimensions (Henly, Kallas, Klatt, & Swenson, 2003; McGrath & Tschan, 2004). Events and transitions are core occasions in time in the science of health (Chinn, 2012) often associated with critical life situations (Donaldson & Crowley, 1978) or turning points in health/illness trajectories (Wheaton & Gotlib, 1997). Events and transitions mark time, whether as explanatory factor or outcome itself (cf., Singer & Willett, 2003).

Relationships

Classical scientific theory is concerned with relationships among concepts in an interlocking system of laws, or “nomological network” (Cronbach & Meehl, 1955) that includes observables and supports explication of mechanisms creating links among substantive concepts. The e-science perspective emphasizes description, integration, and prediction particularly in complex systems whose characteristics are discovered using big data (Cios & Nguyen, this volume; Guergachi,

2003). Statements about cause and effect involve both effects of causes (such as the effects of nursing interventions) and causes of effects (such as causes of patient falls), now viewed in light of the potential outcomes framework (Coffman, this volume; Rubin & Zell, this volume). Both the potential outcomes framework (that defines cause at the level of the individual) and the e-science perspective accommodate emphasis on the individual case, which is essential if nursing science is to be consistent with and relevant to practice and provide evidence for personalized care (e.g., Horowitz, Cullen, Abell, & Christian, 2013; Nardini, Annoni, & Schiavone, 2012).

The value for person-centered nursing care arises from intuitive recognition that health-illness experiences of individual persons and groups are unique. In the past, the idiographic science needed to undergird the practice imperative for person-centered care was generally set aside in favor of research focused on nomothetics (discovery of general laws). Now, it is understood that individual processes (idiographic science) emphasizing unique experiences are more fundamental (Boker, Molenaar, & Nesselroade, 2009; Henly, Wyman, & Findorff, 2011; e.g., Docherty, Sandelowski, & Preisser, 2006). Models for the individual serve as the basis for discovery of more widely applicable “laws” of health, which are linked but not necessarily similar in functional form or parameters (Molenaar, 2004). Understanding both idiographic and nomothetic processes is essential to nursing science.

The research question/the interrogation of data

Framed by theory, gaps in knowledge about health and illness motivate the research question. The nursing research question reflects the subject matter (Grady & McIlvane, Table 1.1, this volume; Hinshaw, Feetham, & Shaver, 1999; National Institute of Nursing Research, 2011), methods of research, and models used to formalize theoretical propositions. Theory becomes established as it withstands repeated encounters with data. Established methods of classical science enable understanding of the extent to which data support or disconfirm theory as represented in a relevant statistical model.

By contrast, the scientific quest in e-science is data-driven (Cios & Nguyen, this volume). The general e-science approach is based on the existence of expansive data. Computationally intensive processes for interrogation of big data are thus necessary and central to advancing knowledge (Bernstam, Smith, & Johnson, 2009). Data-intensive scientific discovery involves machine–human engagement in two basic activities: computing to process data for extraction of patterns, and ascribing meaning to the patterns (regularities) that emerge (Bernstam, Smith, & Johnson, 2009).

Data

Data is the bedrock component of an integrated view of nursing inquiry and research method because scientific theorizing involves observables. In everyday discourse, the term “data” refers to facts arising from observation. As shown in Figure 2.1, the concept *data* (about person and nurse [including their self-care and intervention actions], environment, and health over time) is central. The data concept links the nursing metaparadigm with distinctive ways of theorizing using designed observation, sampling, and statistical models in classical science (branching to the left in Figure 2.1) and data-based explorations in e-science (branching to the right).

Classical science view

In classical science, sample data are “collected,” assembled in a structured dataset, and modeled to obtain conclusions that are generalizable (with some degree of uncertainty) to a defined population. The research question and hypotheses drive procedures used to collect data (make

observations). Data have a priori meaning arising from theory and findings from past research (Polit & Hungler, 1999, Chapters 4–5), with distinctive qualities arising from consideration of measurement error and design. Substantive constructs (concepts) are regarded as extant or hypothetical, sometimes latent, qualities measured fallibly (American Education Research Association et al., 2014; de Vet & Beurskens, this volume; Sijtsma, this volume); this view of data invokes measurement validity and reliability (or precision) as essential to judging the quality of collected data (“Reliability Concepts,” 2015), including data obtained from direct observation of behavior (Harrison, this volume). Item response theory (IRT) extends classical thought about measurement to posit latent respondent traits that are not directly observable to predict responses to items used to elicit trait-relevant information (Baldwin & Wainer, this volume).

Design creates an idealized environment for obtaining data needed to answer a research question within the context of persons, variables, and occasions (Cattell, 1966); structures a dataset; supports a statistical modeling framework; and adds information value that is particularly important when causal hypotheses are entertained (Coffman, this volume; Eckardt & Rindskopf, this volume; Feeley & Cossette, this volume). To determine health effects of self-care and nursing actions, datasets include values for treatment assignment (e.g., intervention or comparator) and treatment assignment mechanism (e.g., randomization or some other balancing mechanism such as matching) that are critical to accurate estimation of a causal estimand (Rubin & Zell, this volume).

e-Science view

In e-science, data are symbols or uninterpreted relational entities (Ackoff, 1989; Graves & Corcoran, 1989) like arrays of numbers. Torrents of health-related data generated from sensors, high-performance computer simulations, high-throughput devices, scientific images, monitoring and therapeutic devices, social media, and electronic and personal health records (D’Avolio, Farwell, & Fiore, 2010; Murdoch & Detsky, 2013; Topol, Schork, & Smith, 2011) are creating a “deluge” (Hey & Trefethen, 2003) that suggests the ubiquity of data, the need to capture (rather than collect) data, and the necessity of curation to make data accessible and useful for scientific purposes (Faniel & Zimmerman, 2011). Big data are characterized by volume (amount, size), veracity (trustworthiness), velocity (speed in appearance and movement), variety (forms; e.g., video, images, text) and value (usefulness) (e.g., Higdon et al., 2013). Big data are structured (coded; Westra, Monsen, & Delaney, this volume) or unstructured (free form, like information in the published literature, narrative nursing notes, or social media posts; see Mudunuri et al., 2013). A wide variety of big data may be combined in single studies (e.g., Abhyankar, Demner-Fushman, Callaghan, & McDonald, 2014).

Although “data” and “information” are sometimes used informally as synonyms, information arises from processed data, refers to “data + meaning,” and requires that data on which it is based be contingently truthful (Floridi, 2005; Staggers & Nelson, this volume). Data are described by metadata (data about data) and provenance (i.e., historical record of its origins and use; Simmhan, Plale, & Gannon, 2005), all of which contribute to information. Information in e-science is akin to data in classical science because both carry meaning.

Variation and covariation/similarity and pattern

Variance refers to differences in some quality judged from a typical tendency; covariance thus suggests multiple qualities with differences from typical tendencies that trend together. Formally, variance refers to the average of the squared deviations from the mean of a numeric variable; and covariance refers to the average cross-product deviation from the means of two variables.

Intraindividual variation (over time or situations) and interindividual variation are distinguished, allowing interindividual differences in intraindividual change to be conceived (Muthén & Muthén, 2000; Nesselroade, 1991; Nesselroade & Rim, 2004) and used as the basis for health trajectory theory (Brant, Beck, & Miaskowski, 2010; Collins, 2006; Henly, Wyman, & Findorff, 2011; Henly, Wyman, & Gaugler, 2011). Comparison of variation within and between groups is the basis of familiar models about equality of means (used in randomized clinical trials and observational studies of treatment effects) and is used in building models for complex, multilevel data.

In e-science, a pattern is “an expression in some language describing a subset of data or a model applicable to the subset” (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Models in this sense capitalize on similarity in features to make a high-level description of data that is valid, novel, useful, and simple. Patterns embodied in big-data-based models become knowledge when they are *interesting* from a domain-specific perspective – valid, novel, simple, and useful for some purpose envisioned by the content expert (user).

Nursing science involves searching for regularities in multiscale features (from molecules to health systems and societies) of observations linked with metaparadigm concepts. Empirico-theoretical structures summarize regularities in the dynamics of health and illness within and across individuals, families, and communities over time and place, including the effects of self-care and nursing interventions. Explaining variation and covariation of variables or co-occurrence of events in light of uncertainty are inherent aspects of classical scientific structures; in nursing science, phenomena are complex, so multivariate, multilevel approaches are ordinarily needed to obtain satisfactory explanations. Identifying hidden patterns based on similarities is fundamental to exploring nursing big data. Explaining variation and covariation via models and uncovering hidden patterns during data exploration generate empirically based theoretical understanding of health outcomes as status, trajectories, or events.

Health: outcomes, status, trajectories, and events

Goals of nursing science involve explaining health and illness processes and creating foundations for nursing and self-care interventions that positively impact health status (Holzemer & Henry, 1999) at one point in time, over a defined period of time, across the lifespan or across the generations. Likewise, self-care and nursing prevention interventions are designed to avoid adverse events that compromise health and contribute to illness and mortality (Sidani & Fleury, this volume), including events constituting errors, slips, and mishaps in provision of health services (Clarke & Schubert, this volume; Institute of Medicine, 2000, 2004). Health indicators in nursing and related health sciences (Bronas & Treat-Jacobson, this volume; de Vet & Beurskens, this volume) have characteristics reflecting a variety of health conceptions (Donaldson & Crowley, 1978; Olebaum, 1974; Smith, 1981), from the simplest classification that judges a person sick or well to graded assessments of symptom intensity to complex depictions of health and illness over time linked with significant health events (Liu & Liu, this volume; Stoddard, this volume).

Models

In classical science, models are representations of associations, mechanisms, and dynamics in theoretical systems (Guergachi, 2003); expression of theoretical models as analogous statistical models that can be estimated and tested is a basic scientific activity (Spanos & McGuirk, 2001). “Model” is directly linked with many other concepts in Figure 2.1. Substantive theory is encapsulated in models – sometimes depicted as path diagrams (e.g., Curran & Bauer, 2007) – that account for variation and covariation in theory-relevant datasets (Jöreskog, 1993). Sample observations arise

via some data-generating mechanism comprising substantive and statistical information (Spanos, 2006, p. 99). The complex array of discrete, countable, ordered, continuous, and trajectory variables may be latent or manifest; understanding the nature of the health indicator is essential for correct specification of models for health phenomena, including response processes (cf., Skrondal & Rabe-Hesketh, 2004, Chapter 2).

Model components

Models are composed of deterministic (fixed) and error components. Theory about substantive concepts and relationships is reflected in the deterministic part; theory about uncertainty reflecting natural variation, measurement error, or unconsidered covariates is reflected in the error part (Skrondal & Rabe-Hesketh, 2004, Section 8.2; also see Lehmann, 2008).

Deterministic functions take many forms that reflect relationships among concepts describing persons, environments, health, nursing, and time. Linear functions are often reasonable representations of relationships between indicators and factors and among constructs in structural equation models (Schmiege & Bryan, this volume) and in multilevel models (Grittner & Lahmann, this volume). Item response theory models use nonlinear functions to link level of a trait being measured with probability of selecting particular response option on a self-report survey (Baldwin & Wainer, this volume). Time as predictor involves linear or nonlinear functional forms for modeling health trajectories (Brant, Beck, & Miaskowski, 2010; Cudeck & Haring, 2007; Henly, Wyman, & Findorff, 2011; e.g., Stoddard, this volume). Modeling time to a health event introduces unique challenges associated with censored observations, but affords opportunities for linking trajectories of clinically relevant health indicators with critical health events such as mortality (Liu & Liu, this volume). Modeling dynamical systems to identify regulatory parameters in situations like patient-nurse interactions or effectiveness of time-based treatments is based on differential equation models (Deboeck & Boker, this volume). Relationships among biophysical concepts likewise involve the dynamics of complex, changing systems (Bronas & Treat-Jacobson, this volume).

It is essential for the probabilistic component of the model to accurately reflect the nature of the outcome (as discrete, graded, continuous, etc.) so that response processes are accurately reflected. The key points here are that a scientific nursing theory involves observables, and complete specification of theory includes assumptions about uncertainty that are consistent with qualities of the observables as classifications, measurements, or counts.

Populations and samples

Inferential statistical procedures involve using observations in a sample to come to conclusions about characteristics of a population; Cochran (1977) details sampling techniques. Populations may be hierarchical (nested; see Grittner & Lahmann, this volume). When populations are hierarchical, (including cluster randomized trials), the number of clusters, the number of cases within clusters, and within-cluster correlation all influence sampling and inference (Bingenheimer & Raudenbush, 2004). In longitudinal studies with occasions nested within persons, duration and frequency of observation, number of persons and attrition, and mechanisms of missingness influence inference (Raudenbush & Xiao-Feng, 2001; Yang & Maxwell, 2014).

Theory is mapped onto statistical models with unknown model parameters that are “constants” governing essential relationships summarized in the model. Small or large sample data are used to make inferences about population parameters with a specified degree of precision using point estimates with standard errors or interval estimates at some level of confidence (Lehmann &

Casella, 1998). Frequentist and Bayesian schools of thought are used to support statistical decision-making (Kass, 2011). The logic underlying these perspectives for inferring causality in experiments and observational (quasi-experimental) study of treatment effects is outlined in Rubin and Zell (this volume). (Also see Coffman, this volume; Eckardt & Rindskopf, this volume.) Rationale for testing, selecting, and generating models in multivariate covariation analysis (e.g., factor analysis, structural equation modeling) is outlined in Schmiede and Bryan (this volume).

Explorations

The heart of e-science is big data, and the overarching goal is to make sense of it (Cios & Nguyen, this volume). A question may be posed (Herland, Khoshgoftaar, & Wald, 2014), but e-science methods are fundamentally oriented toward discovery of hidden patterns – based on similarity of features – in massive data. Theoretical explanation as ordinarily understood is not a primary goal. Findings may serve to generate hypotheses (Berger & Berger, 2004); or, following data integration and description, may be used for creation of actionable knowledge and decision support. The notion of actionable knowledge arising from prediction models hearkens back to the idea of practice theory (Dickoff, James, & Wiedenbach, 1968) because it incorporates goals (reflecting values) and prescriptions, showing how various situations may be produced.

Much big data – like ecological momentary assessment, real-time telemetry from critical care units, care documentation in personal and electronic health records, and health-related tweets – arise authentically from settings in which findings will be applied, suggesting a natural generalizability to research results. Human involvement is essential to interpretation of machine learning-based exploration, by use of labels in supervised approaches and by naming recognized findings when unsupervised approaches are used to make sense of data. In simulation situations, data are generated and integrated to model complex, dynamic systems; applications from systems biology like the virtual physiological human initiative of the Physiome Project (<http://vph-portal.eu/home>) have particular relevance for critical care (Schallom, Thimmesch, & Pierce, 2011) and other areas of nursing and nursing science (Founds, 2009).

Emerging and priority areas of nursing science increasingly involve big data (Conley, this volume; Genomic Nursing State of the Science Advisory Panel, 2013; Grady, 2014; Henly et al., in press). Like other health sciences, nursing science-relevant big data include data on spatiotemporal scales from molecules and tissues/organs (Greve et al., this volume), to patients/persons, and populations; the level at which data are generated and at which questions are posed are relevant to selection of methods in e-science research (Herland, Khoshgoftaar, & Wald, 2014).

Methods for advancing e-science are advancing at breathtaking speeds, reflecting the recognized potential and the youth of the field (<http://bd2k.nih.gov>). Computer science, statistics, and information science underlie methods including analytics (data mining, machine learning), visualization (Cios & Nguyen, this volume), and simulation (Yang, 2010). Big data approaches are only now beginning to be seen in the mainstream of nursing science. Theorizing using e-science methods and findings has not been systematically addressed.

Translation to practice

Better patient care has long been the *raison d'être* for nursing research (Abdellah & Levine, 1965). The wider knowledgescape of the discipline is enhanced when findings from quantitative research are combined with insights obtained from qualitative studies (see Beck, 2013), linked through mixed qualitative and quantitative methods (Doyle, this volume), and reaching consilience in integration with practice utility and ethics (Doane & Varcoe, 2005; Liaschenko & Fisher, 1999).

Dissemination (Dougherty & Jairath, this volume), comparative effectiveness research (Cullum & Dumville, this volume) including economic evaluation (Uchida-Nakakoji & Stone, this volume), and evaluation of nursing systems for safe, quality practice (Clarke & Schubert, this volume) are prerequisites to translating scientific findings into practice – which is a science unto itself (ncats.nih.gov). The vision of personalized therapeutics, initiated by discoveries in genomics for disease treatment (Hamburg & Collins, 2010), extends to personalized behavioral treatments enabled by self-quantification (Swan, 2013).

Conclusion

Theory and theoretical thinking are highly valued in nursing science. Theorizing in contemporary nursing science involves classical statistical reasoning augmented by knowledge discovery processes for big data research. Development of idiographic theory aimed at understanding and predicting course and outcomes for personalized care of specific individuals, families and communities is a new challenge for the 21st century.

References

- Abdellah, F.G., & Levine, E. (1965). *Better patient care through nursing research*. London: Macmillan.
- Abhyankar, S., Demner-Fushman, D., Callaghan, F.M., & McDonald, C. J. (2014). Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. *Journal of the American Medical Informatics Association*, 21, 801–807.
- Ackoff, R.L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16, 3–9.
- American Education Research Association [AERA], American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Beck, C. T. (Ed.). (2013). *Routledge international handbook of qualitative nursing research*. London: Routledge/Taylor & Francis.
- Berger, A., & Berger, C. (2004). Data mining as a tool for research and knowledge development in nursing. *CIN: Computers, Informatics, Nursing*, 22, 123–131.
- Bernstam, E.V., Smith, J. W., & Johnson, T.R. (2009). What is biomedical informatics? *Journal of Biomedical Informatics*, 43, 104–111.
- Bingenheimer, J. B., & Raudenbush, S.W. (2004). Statistical and substantive inferences in public health: Issues in the application of multilevel models. *Annual Review of Public Health*, 25, 53–77.
- Blum, R.L. (2014). The (dis)unity of nursing science. *Nursing Philosophy*, 14, 250–260.
- Boker, S.M., Molenaar, P.C.M., & Nesselroade, J. R. (2009). Issues in intraindividual variability: Individual differences in equilibria and dynamics. *Psychology and Aging*, 24, 858–862.
- Brant, J.M., Beck, S., & Miaskowski, C. (2010). Building dynamic models and theories to advance the science of symptom management research. *Journal of Advanced Nursing*, 66, 228–240.
- Cairns, R. B., Elder, G. H., Jr., & Costello, E. J. (Eds.) (1996). *Developmental science*. Cambridge: Cambridge University Press.
- Cattell, R. B. (1966). The data box: It's ordering of total resources in terms of possible relational systems. In *Handbook of multivariate experimental psychology* (pp. 67–128). Chicago, IL: Rand McNally.
- Chinn, P.L. (Ed.) (2012). Transitions [Special issue]. *Advances in Nursing Science*, 35(3).
- Chinn, P.L., & Kramer, M. K. (2011). *Integrated theory and knowledge development in nursing* (8th ed.). St. Louis, MO: Elsevier Mosby.
- Cochrane, W.K. (1977). *Sampling techniques* (3rd ed.). New York, NY: Wiley.
- Cohen, J. E. (2004). Mathematics is biology's new microscope; biology is mathematics' new physics. *PLoS Biology*, 2(12), e439.
- Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, 57, 505–528.
- Cronbach, L. J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Cudeck, R., & Harring, R. L. (2007). Analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology*, 58, 615–637.

- Curran, P. J., & Bauer, D. J. (2007). Building path diagrams for multilevel models. *Psychological Methods*, 12, 283–297.
- D’Avolio, L. W., Farwell, W. R., & Fiore, L. D. (2010). Comparative effectiveness research and medical informatics. *American Journal of Medicine*, 123 (12 Suppl. 1), e32–e37.
- Dada, J. O., & Mendes, P. (2011). Multi-scale modelling and simulation in systems biology. *Integrative Biology*, 3, 86–96.
- Dickoff, J., James, P., & Wiedenbach, E. (1968). Theory in a practice discipline. Part I: Practice oriented theory. *Nursing Research*, 17, 415–435.
- Doane, G. H., & Varcoe, C. (2005). Toward compassionate action. Pragmatism and the inseparability of theory/practice. *Advances in Nursing Science*, 28(1), 81–90.
- Docherty, S. L., Sandelowski, M., & Preisser, J. S. (2006). Three months in the symptom life of a teenage girl undergoing treatment for cancer. *Research in Nursing and Health*, 29, 294–310.
- Donaldson, S. K. (2003). Commentary. The science of human health – Domain structure. *Biological Research for Nursing*, 4, 165–169.
- Donaldson, S. K., & Crowley, D. (1978). The discipline of nursing. *Nursing Outlook*, 26, 113–120.
- Faniel, I. M., & Zimmerman, A. (2011). Beyond the data deluge: A research agenda for large-scale data sharing and reuse. *International Journal of Digital Curation*, 6, 58–69.
- Fawcett, J. (1984). The metaparadigm of nursing: Present status and future refinements. *Image – The Journal of Nursing Scholarship*, 16, 84–87.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37–54.
- Floridi, L. (2005). Is semantic information meaningful data? *Philosophy and Phenomenological Research*, 70, 351–370.
- Found, S. A. (2009). Introducing systems biology for nursing science. *Biological Research for Nursing*, 11, 73–80.
- Genomic Nursing State of the Science Advisory Panel, Calzone, K. A., Jenkins, J., Bakos, A. D., Cashion, A. K., Donaldson, N., . . . Webb, J. A. (2013). A blueprint for genomic nursing science. *Journal of Nursing Scholarship*, 45, 96–104.
- Gortner, S. R. (1983). The history and philosophy of nursing science and research. *Advances in Nursing Science*, 5(2), 1–8.
- Gottlieb, N. L., & Feeley, N. (1995). Nursing intervention studies. Issues related to time and timing in children and families. *Canadian Journal of Nursing Research*, 27, 13–29.
- Grady, P. A. (2010). Translational research and nursing science. *Nursing Outlook*, 58, 164–166.
- Grady, P. A. (2014, September). *NINR: Advancing innovations for complex health issues*. Paper presented at the Council for the Advancement of Nursing Science State of the Congress, Washington, DC.
- Graves, J. R., & Corcoran, S. (1989). The study of nursing informatics. *Image – The Journal of Nursing Scholarship*, 21, 227–231.
- Guergachi, A. A. (2003). Connecting traditional sciences with the OLAP and data mining paradigms. In V. Belur & V. Dasarathy (Eds.), *Proceedings of SPIE, 5098, Data Mining and Knowledge Discovery: Theory, Tools, and Technology V*, 226–234.
- Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine [Editorial]. *New England Journal of Medicine*, 363, 301–304.
- Henly, S. J. (2007). Lost in time. The person in nursing research [Editorial]. *Nursing Research*, 56, 147.
- Henly, S. J., Kallas, K. D., Klatt, C. M., & Swenson, K. K. (2003). The notion of time in symptom experiences. *Nursing Research*, 52, 410–417.
- Henly, S. J., McCarthy, D. O., Wyman, J. F., Stone, P. W., Redeker, N. S., McCarthy, A. M., . . . The Council for the Advancement of Nursing Science Idea Festival Advisory Committee. (in press). Nursing science: Emerging priorities and implications for nursing scientist education. *Nursing Outlook*, 63.
- Henly, S. J., Wyman, J. F., & Findorff, M. J. (2011). Health and illness over time: The trajectory perspective in nursing science. *Nursing Research*, 60(Suppl. 3), S5–S14.
- Henly, S. J., Wyman, J. F., & Gaugler, J. E. (2011). Health trajectory research: A call to action for nursing science. *Nursing Research*, 60(Suppl. 3), S79–S82.
- Herland, M., Khoshgofaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big Data*, 1(2).
- Hey, T., & Trefethen, A. (2003). The data deluge: An e-science perspective. In F. Berman, G. C. Fox, & A. J. G. Hey (Eds.), *Grid computing: Making the global infrastructure a reality* (pp. 809–824). Chichester: Wiley.
- Higdon, R., Haynes, W., Stanberry, L., Stewart, E., Yandl, G. Howard, C., . . . Kolker, E. (2013). Unraveling the complexities of life sciences data. *Big Data*, 1, 42–49.

- Hinshaw, A. S., Feetham, S. L., & Shaver, J. L. F. (Eds.). (1999). *Handbook of clinical nursing research*. Thousand Oaks, CA: Sage.
- Hoffmann, R. (2003). Why buy that theory? *American Scientist*, 91, 9–11.
- Holzemer, W. L., & Henry, S. B. (1999). Therapeutic outcomes sensitive to nursing. In A. S. Hinshaw, S. L. Feetham, & J. L. F. Shaver (Eds.), *Handbook of clinical nursing research* (pp. 185–198). Thousand Oaks, CA: Sage.
- Horowitz, R. I., Cullen, M. R., Abell, J., & Christian, J. B. (2013). (De)personalized medicine. *Science*, 339, 2255–2256.
- Hyman, R., & Corbin, J. (Eds.) (2001). *Chronic illness: Research and theory for nursing practice*. New York, NY: Springer.
- Institute of Medicine. (2000). *To err is human: Building a safer health system*. Washington, DC: National Academies Press.
- Institute of Medicine. (2004). *Keeping patients safe: Transforming the work environment of nurses*. Washington, DC: National Academies Press.
- Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Thousand Oaks, CA: Sage.
- Kass, R. E. (2011). Statistical inference: The big picture. *Statistical Science*, 26, 1–9.
- Krieger, N. (2001). Theories for social epidemiology in the 21st century: An ecosocial perspective. *International Journal of Epidemiology*, 30, 668–677.
- Lehmann, E. L. (2008). On the history and use of some standard statistical models. *IMS Collections: Probability and Statistics: Essays in Honor of David A. Freeman*, 2, 114–126.
- Lehmann, E. H., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York, NY: Springer.
- Liaschenko, J., & Fisher, A. (1999). Theorizing the knowledge that nurses use in the conduct of their work. *Scholarly Inquiry for Nursing Practice*, 13, 29–41.
- Louis, G. M. B., & Sundaram, R. (2012). Exposome: Time for transformative research. *Statistics in Medicine*, 31, 2569–2575.
- Marriner-Tomey, A., & Alligood, M. R. (2002). *Nursing theorists and their work* (5th ed.). St. Louis, MO: Mosby.
- McGrath, J. E., & Tschan, F. (2004). *Temporal matters in social psychology*. Washington, DC: American Psychological Association.
- Melnik, B. M., & Morrison-Beedy, D. (2012). *Intervention research. Designing, conducting, analyzing, and funding*. New York, NY: Springer.
- Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives*, 2, 201–218.
- Mudunuri, U. S., Khouja, M., Repetski, S., Venkataraman, G., Che, A., Luke, B. T., . . . Stephens, R. M. (2013). Knowledge and theme discovery across very large biological data sets using distributed queries: A prototype combining unstructured and structured data. *PLOS ONE*, 8(12), e80503.
- Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA*, 309, 1351–1352.
- Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, 24, 882–891.
- Nardini, C., Annoni, M., & Schiavone, G. (2012). Mechanistic understanding in clinical practice: Complementing evidence-based medicine with personalized medicine. *Journal of Evaluation in Clinical Practice*, 18, 1000–1005.
- National Institute of Nursing Research [NINR]. (2011). *NINR strategic plan: Bringing science to life*. (NIH Publication 11–7783). Bethesda, MD: National Institutes of Health.
- Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 92–105). Washington, DC: American Psychological Association.
- Nesselroade, J. R., & Rim, N. (2004). Studying intraindividual variability: What we have learned that will help us understand lives in context. *Research in Human Development*, 1(1 & 2), 9–29.
- Nicholl, L. H. (Ed.). (1986). *Perspectives on nursing theory*. Glenview, IL: Scott Foresman.
- Nightingale, F. (1860). *Notes on nursing: What it is and what it is not*. New York, NY: D. Appleton. (Republished in 1969, New York, NY: Dover)
- Novak, J. D., & Cañas, A. J. (2008). *The theory underlying concept maps and how to construct them*. Technical Report IHMC Cmap Tools 2006–01 Rev 01–2008. Pensacola, FL: Florida Institute for Human and Machine Cognition.

- Olebaun, C. H. (1974). Hallmarks of adult wellness. *American Journal of Nursing*, 74, 1623–1625.
- Polit, D. F., & Hungler, B. P. (1999). *Nursing research. Principles and methods* (6th ed.). Philadelphia, PA: Lippincott.
- President's Information Technology Advisory Committee. (2005). *Computational science: Ensuring America's competitiveness*. Arlington, VA: National Coordination Office for Information Technology Research and Development.
- Rapaport, S. M. (2012). Biomarkers intersect with the exposome. *Biomarkers*, 17, 483–489.
- Raudenbush, S., & Xiao-Feng, L. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, 6, 387–401.
- Reliability Concepts and Methods: Special Focus Section. (2015). *Nursing Research*, 64.
- Risjord, M. (2010). *Nursing knowledge: Science, practice, and philosophy*. Chichester: Wiley-Blackwell.
- Schallom, L., Thimmesch, A. R., & Pierce, J. D. (2011). Systems biology in critical care nursing. *Dimensions of Critical Care Nursing*, 30, 1–7.
- Shaver, J. F. (1985). A biopsychosocial view of human health. *Nursing Outlook*, 42, 15–25.
- Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *ACM SIGMOD*, 34(3), 31–36.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis. Modeling change and event occurrence*. New York, NY: Oxford University Press.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling. Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.
- Smith, J. A. (1981). The idea of health: A philosophical inquiry. *Advances in Nursing Science*, 3(3), 43–50.
- Smith, M. J., & Liehr, P. R. (2008). *Middle range theory for nursing* (2nd ed.). New York, NY: Springer.
- Spanos, A. (2006). Where do statistical models come from? Revisiting the problem of specification. *IMS Lecture Notes – Monograph Series. 2nd Lehmann Symposium – Optimality*, 49, 98–119.
- Spanos, A., & McGuirk, A. (2001). The model specification problem from a probabilistic reduction perspective. *American Journal of Agricultural Economics*, 83, 1168–1176.
- Swan, M. (2013). The quantified self. Fundamental disruption in big data science and biological discovery. *Big Data*, 1(2), 85–99.
- Topol, E. J., Schork, N. J., & Smith, J. M. (2011). Digital medicine and the Scripps Translational Science Institute. *Clinical and Translational Science*, 4, 8–9.
- Webster, M. M., & Ward, A. J. W. (2011). Personality and social context. *Biological Reviews*, 86, 759–773.
- Wheaton, B., & Gotlib, I. H. (1997). Trajectories and turning points over the life course: Concepts and themes. In I. H. Gotlib & B. Wheaton (Eds.), *Stress and adversity over the life course. Trajectories and turning points* (pp. 1–28). New York, NY: Cambridge University Press.
- Yang, M., & Maxwell, S. (2014). Treatment effects in randomized longitudinal trials with different types of nonignorable dropout. *Psychological Methods*, 19, 188–210.
- Yang, Z. R. (2010). *Machine learning approaches to bioinformatics*. Hackensack, NJ: World Scientific.
- Yao, Y. Y. (2004). Concept formation and learning: A cognitive informatics perspective. In C. Chang, W. Kinsner, Y. Wang, & D. M. Miller (Eds.), *Proceedings of the 3rd IEEE International Conference on Cognitive Informatics* (pp. 42–51). Retrieved from <http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=9252>

PART II

Measurement

This page intentionally left blank

3

CLASSICAL TEST THEORY

Klaas Sijtsma

This chapter discusses the classical test model, which splits the test score into the sum of the true score and a random measurement error. The two basic concepts in classical test theory (CTT) are reliability and validity. I discuss the definition of test-score reliability, and I notice that test-score reliability is a group characteristic that should be distinguished from the measurement precision of individual test scores. I discuss several well-known methods for estimating test-score reliability. Next, I discuss the concept of measurement precision, and I explain how to statistically test whether a given test score differs from a cut score, from another person's test score, and whether a person's posttest score differs from his/her pretest score so as to establish whether a treatment was successful after a specified, theoretically justified period of time. Finally, I discuss that validity refers to the attribute the test measures but also to the purposes for which a test score can be used. I argue that both aspects of validity are important and that one cannot be ignored without impairing the other.

Test scores, true scores, and measurement error

A psychological test is a systematic classification or measurement procedure that allows statements about theoretically and empirically founded attributes such as abilities and traits, by objectively analyzing the tested individual's responses to a set of standardized items and comparing the results to those of other tested individuals (Drenth & Sijtsma, 2006, p. 67). Examples of abilities from the intelligence domain are spatial orientation and transitive reasoning, and examples of traits are introversion and neuroticism. The health and nursing sciences study measurement of attributes such as pain experienced by patients suffering from burn wounds (De Jong, Bremer, Schouten, Tuinebreijer, & Faber, 2005), health-related quality of life aspects such as physical functioning, general health perceptions, vitality, and social functioning (Gandek, Sinclair, Kosonski, & Ware, 2004), and adherence to medication and lifestyle for patients with hypertension (Ma, Chen, You, Luo, & Xing, 2012). The health and nursing sciences have adopted psychometrics for measuring the typical attributes using multi-item measurement instruments, where items can be printed rating-scale statements (questionnaire), oral (interview) questions, or structured observations.

CTT is a psychometric method for constructing tests and analyzing test performance. I introduce some notation before I discuss the basic idea of CTT. Let a test consist of J items, and let X_j be the random variable and score x_j on item j ($j = 1, \dots, J$) the realization. For simplicity, I

assume that different items have the same response format and are scored similarly; hence, scores on item j are denoted $x_j = 0, \dots, m$. The test score X_+ is the sum of the J item scores,

$$X_+ = \sum_{j=1}^J X_j.$$

Index ν enumerates persons, so that $X_{+\nu}$ represents the test score of person ν .

The basic idea of CTT is that measurement suffers from random measurement error. Consequently, measurement error distorts a person's test performance by an unknown magnitude and in an unknown direction. Measurement error may be due to unpredictable variation in a person's mood, concentration, attention span, susceptibility to external stimuli, and physical condition expressed in general feelings of well-being or lack thereof (e.g., susceptibility to headache). In a thought experiment where the same person responds to test items repeatedly under the same conditions, due to random measurement error one would obtain a distribution of test scores $x_{+\nu}$ rather than one fixed value. Lord and Novick (1968, p. 30) called this hypothetical distribution across replications the propensity distribution for each person ν .

Definition: The true score of person ν is the mean of his/her propensity distribution, and technically is the expected test score,

$$T_\nu = \varepsilon(X_{+\nu}).$$

Definition: Random measurement error is the difference between a test score obtained on one testing occasion (i.e., a draw from the propensity distribution) and the true score,

$$E_\nu = X_{+\nu} - T_\nu.$$

Propensity distributions of different persons have different true scores and standard deviations; see Figure 3.1.

By definition the errors for one person cancel when averaged, so that

$$\varepsilon(E_\nu) = 0.$$

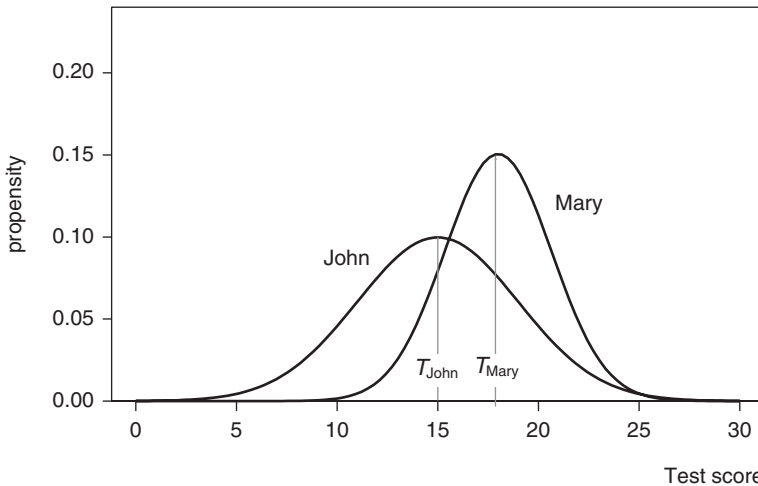


Figure 3.1 Two propensity distributions with different means (true scores) and standard deviations.

Definition: Based on the definitions of true score and measurement error the classical test model is defined as

$$X_{+v} = T_v + E_v.$$

CTT concentrates on the person's true score, which represents his/her average test performance expressing what the person can accomplish (maximum achievement) or do (typical performance) without the distortion induced by random measurement error in one particular test administration. In practice, tests are constructed so as to reduce the influence of measurement error.

As individual propensity distributions are theoretical and thus unavailable in practical research and clinical testing situations, I switch to realistic, single test administration and consider the test scores obtained in a group of persons. So far, I related measurement error to test scores but CTT assumes that measurement error occurs with any measurement value including those obtained for specific test items; hence, item scores can also be split into true item scores and random measurement errors. It follows that in a group of persons measurement errors on different items j and k are uncorrelated; that is, let ρ denote the product-moment correlation, then $\rho_{E_j E_k} = 0$.

Next, I introduce the definition of parallel tests. Parallel tests are different but interchangeable versions of the same test that replace the hypothetical replications of the same test as they appear in the propensity distribution. Let the test scores on two parallel tests be denoted X_+ and X'_+ (and likewise, T and T' and E and E').

Definition: Two tests are parallel if

$$T_v = T'_v, \text{ for all } v;$$

and, letting σ^2 denote the variance, in the group,

$$\sigma_{X_+}^2 = \sigma_{X'_+}^2;$$

This definition can be shown to imply that parallel tests have exactly the same psychometric properties in a group of persons. However, due to random measurement error an individual is expected to have different observable test scores on the parallel tests; that is, $E_v \neq E'_v$, implying $X_{+v} \neq X'_{+v}$.

Definition: Reliability is the correlation between the test scores on two parallel tests in the group under consideration, and is denoted $\rho_{X_+ X'_+}$.

As the parallel tests are interchangeable by definition, the correlation $\rho_{X_+ X'_+}$ gives the reliability of each separate test. Note that in the unrealistic situation where there is no measurement error, $\rho_{X_+ X'_+} = \rho_{TT'} = 1$, which is the maximum reliability value possible. Since two parallel test scores can be considered independent draws from the group members' propensity distributions, the correlation $\rho_{X_+ X'_+}$ expresses the degree to which random measurement error distorts the equality of the two sets of measurements. The more the draws are alike, the closer the reliability is to its maximum value 1.

Measurement errors are uncorrelated with the true scores from the same test (i.e., $\rho_{ET} = 0$). The zero-correlation implies that test-score variance is the sum of true-score variance and error variance: $\sigma_{X_+}^2 = \sigma_T^2 + \sigma_E^2$. Using this result, it can be shown that, for test score X_+ , reliability can be written in three different ways,

$$\rho_{X_+ X'_+} = \frac{\sigma_T^2}{\sigma_{X_+}^2} = 1 - \frac{\sigma_E^2}{\sigma_{X_+}^2}. \quad (3.1)$$

Hence, reliability increases as parallel test scores are more alike (left), true-score variance relative to test-score variance is larger (middle), and error-score variance relative to test-score variance is smaller (right). The second term in Equation 3.1 shows that the reliability has a minimum equal to 0 (when $\sigma_T^2 = 0$) and a maximum equal to 1 (when $\sigma_T^2 = \sigma_{X_+}^2$); hence, $0 \leq \rho_{X_+X'_+} \leq 1$. Equivalently, the third term in Equation 3.1 shows that reliability equal to 0 corresponds to $\sigma_E^2 = \sigma_{X_+}^2$ and reliability equal to 1 corresponds to $\sigma_E^2 = 0$. The definitions and interpretations of the reliability suggest two important results that often are not well understood.

Result: Test-score reliability depends on the population, and the same test usually has different reliability values in different populations.

Comment: If a test has reliability equal to, say, .91 in a group of 6–12-year-olds, this value cannot be assumed to be valid in the separate age groups as well. Instead, if the test is used in particular age groups, the reliability has to be determined for each group separately. Also, if a test has a reliability of, say, .87 in the group of 8-year-olds, one cannot simply assume that it is also reliable in the other age groups. Test-score reliability has to be established for each group separately.

Result: Test-score reliability provides information about the variation between true scores in a group relative to the test-score variation; hence, test-score reliability provides between-persons information. Reliability does not provide within-persons information like the person's propensity distribution does, and thus it is uninformative about measurement precision for individuals.

Comment: Conclusions such as “The test has reliability equal to .82, which is high enough” (compared to an accepted rule of thumb, such as .80) do not help to decide whether a particular individual's test score is significantly higher than a cut score used for selection and rejection or whether a particular change score indicates significant improvement due to therapy. To make such decisions about individual test scores, one needs to consider measurement precision rather than reliability (Mellenbergh, 1996).

Next, I discuss methods for estimating reliability followed by the estimation of measurement precision for different measurement problems concerning individual measurement.

Estimating reliability

Methods based on single test administration

Coefficient alpha and higher lower bounds

Coefficient alpha is the most frequently used method for reliability estimation. Let σ_j^2 be the variance of item score X_j , σ_{jk} the covariance between X_j and X_k , and $\bar{\sigma}$ the mean of the inter-item covariances. I provide three different equations for coefficient alpha (Cronbach, 1951; Guttman, 1945) that are mathematically equivalent. The first equation (Equation 3.2) is the best known, then the second (Equation 3.3), and the third equation (Equation 3.4) is the least known but the most informative about what alpha actually is:

$$\alpha = \frac{J}{J-1} \left(1 - \frac{\sum_{j=1}^J \sigma_j^2}{\sigma_{X_+}^2} \right); \quad (3.2)$$

$$\alpha = \frac{J-1}{J} \left(\frac{\sum \sum_{j \neq k} \sigma_{jk}}{\sigma_{X_+}^2} \right); \quad (3.3)$$

Table 3.1 Example of a covariance matrix for four items

	1	2	3	4
1	.25	.12	.16	.10
2	.12	.24	.08	.09
3	.16	.08	.25	.12
4	.10	.09	.12	.21

and

$$\alpha = \frac{J^2 \bar{\sigma}}{\sigma_{X_+}^2}. \quad (3.4)$$

A computational example is the following. Table 3.1 shows a covariance matrix (Sijtsma, 2009a) for four items that has the item variances σ_j^2 on the main diagonal and the inter-item covariances σ_{jk} on the off-diagonal positions. The statistics needed to compute alpha are $\sum \sigma_j^2 = 0.95$ (Equation 3.2), $\sum \sum \sigma_{jk} = 1.34$ (Equation 3.3), $\bar{\sigma} = 1.34 / 12 \approx 0.112$ (Equation 3.4), and $\sigma_{X_+}^2 = 0.95 + 1.34 = 2.29$ (all three equations); insertion of these results in the appropriate equation yields $\alpha \approx .78$.

Alpha is a lower bound to the reliability, which means that alpha is at most as high as the reliability of the test score. The theorem that formalizes this property is

$$\alpha \leq \rho_{X_+X'_+};$$

see Novick and Lewis (1967) for the proof. The theorem posits that if in a population a test score has an alpha value of, say, .78, the true reliability is at least equal to .78; that is, $\rho_{X_+X'_+} \geq .78$. Novick and Lewis (1967) showed that $\alpha = \rho_{X_+X'_+}$ if and only if all J items in the test are essentially τ -equivalent (where $\tau = T$); this is an equivalence condition for the items that is somewhat weaker than parallelism but still unrealistically restrictive. For example, essential τ -equivalence implies that the inter-item covariances for all the item pairs in the test are equal. In real data, inter-item covariances always show variation. Consequently, for real tests alpha is always lower than the reliability,

$$\alpha < \rho_{X_+X'_+}.$$

The results for alpha are theoretical; due to sampling fluctuation estimates obtained in samples can be higher than the unknown population test-score reliability. Thus, in real research one cannot simply say that $\alpha = .78$ implies $\rho_{X_+X'_+} \geq .78$. Several authors have studied the sampling properties of alpha and provide methods to estimate a confidence interval (CI) for alpha that shows the uncertainty in the value found due to the sample size used (Van der Ark, Van der Palm, & Sijtsma, 2011).

Despite its popularity, coefficient alpha is one of the worst methods for reliability estimation. Bentler and Woodward (1980) and Ten Berge, Snijders, and Zegers (1981) studied the problem of finding the greatest lower bound (glb) to the reliability based on one test administration. The problem they solved was that of finding the lowest possible reliability given the data and the CTT assumption of uncorrelated measurement errors. The outcome of this optimization problem, called the glb, is not based on a simple equation (hence, I cannot provide one), and has the mathematical property that

$$\alpha \leq \text{glb} \leq \rho_{X_+X'_+},$$

with equalities if and only if all items are essentially τ -equivalent. In real test data, strict inequalities are realistic. Then we have the following intriguing result.

Result. Coefficient alpha can only attain values that the test-score reliability cannot have in the population for the test under consideration.

Comment. If the glb equals, say, .88, then it follows that $[\text{.88} \leq \rho_{X_+X'_+} \leq 1]$. Alpha is always smaller than the glb and thus lies below the interval of reliability values that are consistent with the particular data set. As the reliability lies in the interval $[\text{glb}, 1]$ one cannot rule out that $\rho_{X_+X'_+}$ is very high and even equal to 1. The uncertainty is caused by the availability of scores for one test whereas by definition the reliability requires scores for two parallel tests; thus, half of the data is missing.

The glb is available, for example, in the program EQS – Structural Equation Modeling Software (www.mvsoft.com/) – which also provides coefficient alpha for the purpose of comparison. Guttman (1945) discussed coefficient alpha under the name of lambda3 and also five other reliability methods that are available in SPSS using model = Guttman on the “scale” module. One of the other methods is lambda2, and it can be proven that (e.g., Sijtsma, 2009a; Ten Berge & Sočan, 2004; Ten Berge & Zegers, 1978),

$$\alpha \leq \text{lambda2} \leq \text{glb} \leq \rho_{X_+X'_+}.$$

Thus, although outside the range of admissible reliability values, lambda2 is closer to the glb than alpha, and in my experience it usually is higher than alpha by .01 or .02 units.

Result: Coefficient alpha is a lower bound to the reliability.

Comment: Alpha values are too low to represent values that the reliability $\rho_{X_+X'_+}$ actually could have. Lambda2 is a higher lower bound and the glb is the highest.

Coefficient alpha is often used as an index for internal consistency, with high values of alpha thought to mean that items are, as Cronbach (1951, p. 320) put it, “psychologically interpretable” but not necessarily meaning “that all items be factorially similar.” Conflating high alpha values with unidimensionality in the factorial sense is incorrect, as Sijtsma (2009a) and several other authors (e.g., Ten Berge & Sočan, 2004) have argued.

Let us assume that a set of items is internally consistent if they measure the same ability or trait, so that one factor explains the structure in a set of inter-item correlations to a great extent. To show that there is not a clear-cut relationship between alpha and the factor structure of the items in the test, Table 3.2 shows two possible covariance matrices and the two corresponding correlation matrices in which correlations are obtained by, using $\sigma_j\sigma_k = .25$ based on the covariance matrices, $\rho_{jk} = \sigma_{jk} / (\sigma_j\sigma_k) = 4\sigma_{jk}$. Correlation matrices are then input for factor analysis. The first correlation matrix clearly shows a 2-factor structure with items 1 and 2 correlating .64 and items 3 and 4 correlating .64, and all other correlations equal to .04. The second correlation matrix shows a one-factor structure, with all items correlating .24.

Table 3.2 shows dissimilar inter-item correlation matrices reflecting dissimilar factor structures. Interestingly, the corresponding covariance matrices reveal that the different factor structures produce the same alpha; for both cases $J = 4$, $\bar{\sigma} = .06$, and $\sigma_{X_+}^2 = 1.72$; hence, $\alpha = 56$. The example shows that different covariance structures representing different factorial compositions can have the same mean covariance value, producing the same alpha; hence, alpha is an inadequate source of information about internal consistency.

The widespread belief that the magnitude of alpha says something about internal consistency is a misconception. Low alpha values (e.g., .6) and high alpha values (.9) each may go together with highly different factorial structures; see Sijtsma (2009a) for examples.

Table 3.2 Two covariance matrices and corresponding correlation matrices producing the same alpha

Covariance Matrices									
	1	2	3	4		1	2	3	4
1	.25	.16	.01	.01		.25	.06	.06	.06
2	.16	.25	.01	.01		.06	.25	.06	.06
3	.01	.01	.25	.16		.06	.06	.25	.06
4	.01	.01	.16	.25		.06	.06	.06	.25
Correlation Matrices									
	1	2	3	4		1	2	3	4
1	1	.64	.04	.04		1	.24	.24	.24
2	.64	1	.04	.04		.24	1	.24	.24
3	.04	.04	1	.64		.24	.24	1	.24
4	.04	.04	.64	1		.24	.24	.24	1

Result: Values of alpha are uninformative of the factorial structure of the test.

Comment: I recommend one uses factor analysis or item response theory to study the factorial structure or the dimensionality of the data.

CTT does not restrict the factorial composition of the true score. Hence, the reliability is defined as the correlation between two parallel test scores, irrespective of the test scores' factorial composition. A low reliability does not exclude a one-factor structure for the items and a high reliability does not exclude a multi-factor structure. Reliability simply is not a measure of (one-) factorial inter-item structure.

Split-half reliability

Split-half reliability estimates are sometimes recommended. Here I describe and comment on split-half reliability. The Spearman-Brown (SB) prophecy formula is central to understanding split-half reliability. The SB prophecy formula provides the reliability $\rho_{KK'}$ of a test score after the test was lengthened by a factor K , where each of the K test parts are parallel. The SB formula equals

$$\rho_{KK'} = \frac{K\rho_{X_+X'_+}}{1 + (K-1)\rho_{X_+X'_+}}.$$

The split-half method for estimating reliability entails splitting the test in two half tests that mimic parallel tests and then estimating the correlation between the test scores on the two halves thus obtaining a reliability estimate of a half test. Next, one uses the SB formula with $K = 2$,

$$\rho_{22'} = \frac{2\rho_{X_+X'_+}}{1 + \rho_{X_+X'_+}},$$

to estimate the reliability of a test that contains $J = 2K$ items. The result estimates the reliability well if the two test halves are parallel but in practice this condition is not met.

A test with an even number of items can be divided in two halves in $1/2(J!) / [(J/2)!]^2$ unique ways. For example, with $J = 4$ one has $4!/(2!)^2 = 6$ unique item pairs, and $1/2(4!)/(2!)^2 = 3$ unique

pairings of item pairs: (1,2; 3,4), (1,3; 2,4), and (1,4; 2,3). From among the many possibilities for larger J , it is difficult to choose a division into halves that is both substantively defensible and also yields closely parallel forms. Lord and Novick (1968, p. 93; after Cronbach, 1951) proved the following relationship. For each division, determine coefficient alpha based on the two test halves; using Equation 3.2, denote alpha based on two test halves $\alpha(2)$ and the covariance between the total scores on the two halves σ_{12} , so that, using Equation 3.3,

$$\alpha(2) = \frac{4\sigma_{12}}{\sigma_{x_+}^2}.$$

Denote alpha based on the J items in the test by $\alpha(J)$, as in Equation 3.3. Then, the mean of the $\alpha(2)$ values based on all $1/2(J!)/[(J/2)!]^2$ splits equals the common alpha,

$$\varepsilon[\alpha(2)] = \alpha(J).$$

Lord and Novick (1968, p. 94) noted that, by definition, each $\alpha(2)$ is a lower bound to the reliability. For real data, $\alpha(2)$ values vary relative to their mean, $\alpha(J)$ (i.e., test halves are not essentially τ -equivalent), so that several values are higher lower bounds than $\alpha(J)$. Thus, one could select the highest $\alpha(2)$ value to replace $\alpha(J)$. This is only defensible if the sample is very large so that the $\alpha(2)$ values are estimated with great precision. In smaller, realistic samples this strategy leads to chance capitalization and is not feasible.

Result: To avoid chance capitalization for smaller, realistic sample sizes, split-half reliability is better replaced by a lower bound value based on all J items from a single administration.

Comment: Jackson and Agunwamba (1977) discussed Guttman's lambda4, which is the greatest split coefficient based on all divisions of the test in two parts, either of equal or unequal size. Lambda4 is a lower bound to the reliability. Clearly, it may also suffer from chance capitalization.

Methods based on two test administrations or administration of two tests

Retest reliability

Retest reliability entails that the test is administered twice to the same group of people with a time interval in between that can extend from only a few days to several months or more. The correlation between the test scores on two occasions is considered an estimate of the reliability. Two remarks are in order.

First, as a simulation of parallel tests that form the basis of reliability $\rho_{x_+x'_+}$ the repeated administration of the same test must fail as it is impossible to rule out memory effects on the second test performance. Memory effects suggest that the correlation between the two test scores is higher than $\rho_{x_+x'_+}$.

Second, if the trait one measures changes over time the retest correlation reflects the instability of the trait, showing the influence of people's true-score change on test results in addition to the influence of measurement error. Trait instability suggests a correlation lower than $\rho_{x_+x'_+}$. In practice, retest correlations often are lower than coefficient alpha.

By also expressing trait instability, the retest correlation expresses a characteristic of the trait whereas a reliability estimation method must only express influence of random measurement error upon replication. For example, one may compare this to twice measuring eight-year-olds' body height at a six-month time interval. Assuming that body height can be measured almost without error thus suggesting near perfect reliability, and given that eight-year-olds are expected to gain height at different paces, the retest correlation between height measures is smaller than 1.

But this correlation reflects individual differences in true height change, not influence of measurement error; hence, retest correlation is not only a reliability estimate. (This is only true if the two measures are linearly related but I ignore this subtlety here).

Result: Retest “reliability” provides a contaminated reliability estimate.

Comment: Retest “reliability” picks up trait instability over time. Hence, in addition to random measurement error it also expresses a property of the attribute and thus provides a mixture of reliability and validity information.

Parallel-test reliability

Test constructors rarely assemble two versions of the same test. Not only would it require much effort to construct twice as many high-quality items as one needs for one test but it also remains to be seen whether the resulting tests are parallel. If they are not, their correlation is a biased estimate of test-score reliability. Parallelism can be investigated as follows.

Two parallel test scores have the same correlations with any other variable Y , so that parallel test scores must satisfy

$$\rho_{X_+Y} = \rho_{X'_+Y}, \text{ for all variables } Y. \quad (3.5)$$

The best way to investigate whether two tests are parallel is to examine the correlations each has with any other variable Y ; these correlations must be equal for each variable Y (Equation 3.5). The investigation of all possible variables Y seems to be unnecessary, and may be replaced by a well-chosen, small set.

Result: Parallel-test reliability in many cases is an unpractical method.

Comment: A lower bound method is more convenient and may be recommended.

Different but also more complex approaches to reliability have been proposed in the context of generalizability theory (Brennan, 2001) and structural equation modeling (Raykov & Shrout, 2002). These approaches are rather technical, assume knowledge about analysis of variance and factor analysis, and hence cannot be discussed well in this chapter; also see Sijtsma (2009b) and Sijtsma and Van der Ark (2015) for more discussion.

Measurement precision of individual test scores

Individual test scores are used in clinical practice to (a) aid in making treatment decisions (i.e., judge whether an individual’s test score exceeds a cut score X_c used to make a diagnosis or to determine whether treatment is needed); (b) determine whether the status of two individuals on the same test is different; and (c) assess whether real change in status has occurred on two occasions of measurement (especially, before and after treatment).

For these purposes, one uses the standard error of measurement (SEM). SEM is the standard deviation of the measurement error, and is obtained by rewriting the right-hand side of Equation 3.1,

$$SEM = \sigma_E = \sigma_{X_+} \sqrt{1 - \rho_{X_+X'_+}}.$$

First, one may notice that SEM depends on the standard deviation of the test score and the reliability; thus, SEM is a group characteristic. Second, in making inferences about individual test scores SEM plays the role of the standard error, and is equal for all persons. This means that it is tacitly assumed that propensity distributions have equal spread.

Result: In practical use of individual test scores obtained from CTT-based tests, one assumes that measurement precision expressed by SEM is identical for all persons tested.

Comment: Item response theory is an alternative model (Baldwin & Wainer, this volume; Embretson & Reise, 2000; Sijtsma, Emons, Bouwmeester, Nyklicek, & Roorda, 2008; Van der Linden & Hambleton, 1997) that defines person measurement on a logit scale, and measurement values have a standard error that varies along the scale. Variation depends on the psychometric properties of the items and expresses that the test is a more precise measurement instrument for some persons than it is for others.

SEM is used as follows (e.g., Sijtsma & Emons, 2011). We consider the test score X_+ to be an estimate of the true score T : that is, $\hat{T} = X_+$. Different true-score estimation methods exist but they are more complex than the method I discuss and lead to the same line of reasoning. To test hypotheses about a person's true score, I determine a CI for T . I assume that propensity distributions are normal, such that: $\hat{T}_\nu (= X_{+\nu}) \sim N(T_\nu, \sigma_E^2)$ where σ_E^2 is the squared SEM. Denoting sample variance by S^2 and sample reliability by $r_{X_+X'_+}$, SEM is estimated by

$$S_E = S_{X_+} \sqrt{1 - r_{X_+X'_+}}.$$

Let α be the significance level; then a $(1 - \alpha) \times 100\%$ CI is routinely estimated as

$$[\hat{T}_\nu - z_{\alpha/2} S_E; \hat{T}_\nu + z_{\alpha/2} S_E],$$

where $z_{\alpha/2}$ is the standard normal deviate corresponding to the appropriate area under the normal curve that describes the propensity distribution. Using a table for the standard normal distribution, for a 95% CI one finds that $z_{.025} = 1.96$ and for a 90% CI that $z_{.05} = 1.645$; and so on.

I use the data from Table 3.1, assume that they are sample data, and notice that $S_{X_+} = 1.51$ and $\alpha = .78$; then, $S_E = 1.51 \sqrt{1 - .78} \approx 0.71$. A 95% CI equals $[\hat{T}_\nu - 1.39; \hat{T}_\nu + 1.39]$ and a 90% CI equals $[\hat{T}_\nu - 1.17; \hat{T}_\nu + 1.17]$. Suppose that the items for the Table 3.1 data were 0,1-scored, so that possible test scores are equal to 0, 1, 2, 3, or 4. Propensity distributions in this case are discrete and non-normal, so that the normality assumption used to create the CI is incorrect. For longer tests and for tests consisting of polytomous items, distributions better approximate normality and CIs are more accurate. For the 4-item example, it is particularly important to see that the 95% CI covers 2.78 units of a 4-unit scale and the 90% CI covers 2.34 units. Thus, despite a reliability of .78 the uncertainty about true scores is considerable. This is due to the short length of the scale. To have more certainty, tests must contain more items (Kruyen, Emons, & Sijtsma, 2012; Sijtsma, 2012a). As items are added to a test and the range of possible scores increases, SEM also increases but more slowly than the scale length. Thus, measurement is more precise and CIs cover relatively smaller scale intervals.

The CI can be used to test whether a particular test score is significantly smaller or larger (two-sided test) than a particular cut score that is used to decide, for example, whether the person does or does not need a treatment. For each person tested, one estimates his/her CI and then checks whether the specified cut score X_c is in the interval; if it is not, then the test score is significantly different from the cut score and a reliable decision can be made. If the cut score falls within the CI, then a decision cannot be made, and the practitioner has to decide whether additional information has to be collected. The best policy depends on the problem at hand. One-sided tests can be done by adapting the significance level.

CIs can be used to compare the test scores of two individuals, both of which contain measurement error. For persons ν and w , I ask whether the difference $D_{\nu w} = X_{+\nu} - X_{+w}$ is significantly different from 0; that is, whether under the null hypothesis the true scores are equal: $T_\nu - T_w = 0$. Given the assumptions of CTT, it can be shown that

$$SEM(D) = \sigma_{E(D)} = \sqrt{2}\sigma_E,$$

so that the CI for the true difference denoted Δ_{vw} – estimated by $\hat{\Delta}_{vw} = D_{vw}$ and using the sample estimate of $\sigma_{E(D)}$ denoted $S_{E(D)}$ – equals

$$[\hat{\Delta}_{vw} - z_{\alpha/2}S_{E(D)}; \hat{\Delta}_{vw} + z_{\alpha/2}S_{E(D)}].$$

For a given CI, one checks routinely whether the value 0 lies in the interval. If so, one may assume that $T_v = T_w$; else, one assumes that $T_v \neq T_w$.

Lastly, I consider a person v who was treated after being pretested yielding pretest score $X_{+v,pr}$ and posttested yielding posttest score $X_{+v,po}$, and ask whether the difference $X_{+v,po} - X_{+v,pr}$ is significantly different from 0; that is, whether under the null hypothesis $T_{v,po} - T_{v,pr} = 0$. Several significance tests were proposed but the reliable change (RC) index (Jacobson & Truax, 1991) has the best sampling properties (Atkins, Bedics, McClinchey, & Beauchaine, 2005; Bauer, Lambert, & Nielsen, 2004) and is the one most frequently used in real-change testing across two occasions. RC is defined as

$$RC_v = \frac{X_{+v,po} - X_{+v,pr}}{S_{E(D)}},$$

where $S_{E(D)}$ is the standard error of the difference score under the null hypothesis of no change. It is the same standard error I used for testing whether the true scores of two different persons are equal. RC is assumed to be a standard normal deviate and the usual routine for hypothesis testing can be followed.

Validity of the test and the test score

Validity is a more complex topic than reliability. Reliability is a mathematical concept and even though discussions may be quite involved, they revolve around rather straightforward issues such as how to estimate reliability, what is a sufficiently high value for a particular use of the test, and how reliability should be used. On the other hand, validity involves the issue whether the test measures the intended attribute well, which entails not only the collection of empirical evidence but also eventually the assessment of the evidence through judgment; and validity also involves the empirical investigation of the suitability of the test score for predicting criteria outside the test, such as readiness for a particular treatment. These approaches to validity have been named construct validity and predictive validity, respectively. The distinction between two kinds of validity arose in the 1920s when validity as a concept originated, immediately leading to debate about the issue whether it is the test that is valid, referring to the attribute the test measures, or the test score, referring to the test score's practical uses; see Sireci (2009) for a historical overview.

Present-day validity conceptions predominantly focus on the validity of the test score and the question what the test measures is considered less important. For example, consider Messick's (1989, p. 13) influential proposal that "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores and other modes of assessment." Thus, Messick defined validity as a broad concept referring to the degree to which one can find sources of support for a particular interpretation or a particular use of the test score. This perspective underlies the current *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014). In both cases, the emphasis is on the use of test scores and the question what the test measures moved to

the background. The definition is primarily technical: If the test does what it is supposed to do, it works and one is done. The consequence is that, as long as it works, in principle one does not need to know what the test measures. The question what a test measures does not seem to play a role in these modern views on validity.

The inference from Messick's definition that the question what the test measures is neglected may seem rather extreme, and one may argue against it that a test that is successful at predicting readiness for treatment, for example, must have been constructed well, or otherwise it would not have been a successful predictive instrument. A counterargument is that a criterion like readiness for treatment does not represent an attribute but depends on many different conditions and traits. For example, the treatment may require that people exhibit low levels of anxiety, depression, and neuroticism but also that their social and verbal skills are above a particular level. Obviously, a single test that represents this criterion would measure a hodgepodge of psychological attributes and using a battery of tests each measuring one unique attribute reliably and validly would seem conceptually more appropriate (Nunnally, 1978). This is the stance I take in this chapter: A test must measure one attribute.

Recently, several authors have expressed their concern about the technological approach to validity (Borsboom, Cramer, Kievit, Zand Scholten, & Franić, 2009; Michell, 1999; Sijtsma, 2012b; Zumbo, 2009). The viewpoint they all share is that a theory about the attribute must support the measurement of the attribute. Their concerns can be summarized by the observation that many measurement instruments used in the social and the behavioral sciences lack a sound theoretical basis. When theories are available, the problem usually is that there are too many competing theories for the same attribute, and that no crucial experiments have been done allowing a decision that favors one theory relative to the others. Intelligence is an excellent example. Several theories exist next to one another so that different tests for intelligence can be based on different theories, such as Spearman's two-factor theory, Thurstone's seven-factor theory, and Guilford's three-dimensional 120-factor theory whereas other tests are based on binary distinctions between verbal and performance intelligence and crystallized and fluid intelligence (Carroll, 1993). Conceptualization and measurement of health in health and nursing sciences have challenges similar to intelligence, and many proposals to define health are available in the literature (e.g., Riley, Rothrock, Bruce, Christodolou, Cook, Hahn, & Cella, 2010; Smith, 1981; Ware, 2000). Grady and McIlvane (this volume) identify health as the central concept in nursing science, and the varied notions of health create ongoing challenges to advancement of the science. The richness of the fields in fact signifies their weaknesses as different intelligence and health conceptions continue to exist next to one another. For many personality traits such as leadership and social intelligence the situation is grimmer, as less precise propositions, hypotheses, and guesses replace theories and are often expressed by inaccurate associations between the test score and other variables.

The absence of well-founded and well-tested theories implies that tests often consist of a set of items that define what the test measures. The psychometric analysis of the collected data and the analysis of the correlations of the test score and a limited number of other variables serve as the basis for establishing validity in hindsight rather than a priori based on theory. This approach is known as operationism; the attribute coincides with the operations used to measure it. However, the correct scientific approach is that a theory about the attribute guides the operationalization of the attribute into a set of items that produce the observable responses the theory predicts.

To enhance test validity, a theory about the attribute to be measured must guide the test construction process. When such a theory is unavailable, Borsboom et al. (2009) advise to investigate what persons do when they respond to items: Which cognitive processes are activated? Which affective processes are activated? Psychometrics can lend a helping hand by means of componential processing models (e.g., De Boeck & Wilson, 2004). The linear logistic test model (Fischer,

1974) provides a relatively simple example, which explains item difficulties from the additive contributions of different operations that students have to perform when they attempt to solve a cognitive or other problem. See Jansen and Van der Maas (2002; proportional reasoning) and Smits and De Boeck (2003; guilt) for other examples using psychometric models.

Result: A theory about the attribute must guide the construction of the test.

Comment: My advice to researchers is to use whatever theory about the attribute in question is available to design a first draft of the test. Researchers may use psychometric componential processing models to study the psychological processes that subjects employ to solve the items and to use the results of the statistical data analysis to amend the test and the theory that stood at the basis of the test construction. The flexibility of modern psychometric models including variations on item response theory models, latent class models, and factor models provides good opportunities to find a useful description of the processes and improve the test when necessary. The end result is a test that measures the intended attribute.

After the test has been constructed, it should be investigated how well it can be used for a particular practical purpose, such as a diagnosis of a person or a prediction of the suitability of a person for a particular treatment or a therapy. Here, it is of great importance to meticulously analyze the criterion one wants to predict, so as to understand well how it is composed and which tests and possibly different information sources (e.g., interviews, observations, medical information) one needs to predict accurately. It is important to keep track of the outcomes of the treatment across a longer time period and for a large group of people so as to evaluate how successful the test, the test battery, or the battery of information sources was in making predictions.

Discussion

This chapter provided a critical discussion of CTT. The well-known methods for reliability estimation were discussed and it was concluded that lower bound reliability estimates such as λ^2 and the glb should be preferred over coefficient alpha. In addition, the use of coefficient alpha as an index of internal consistency was discouraged as particular alpha values are uninformative of the test's factorial composition. Factor analysis and item response theory are much better suited to investigate the test's factorial composition or dimensionality, respectively. Finally, it was noticed that the reliability is a group characteristic, expressing in a group of people the correlation between two sets of parallel test scores or test replications or, equivalently, the proportion of test score variance that is true-score variance. For assessing whether an individual test score is significantly different from an error-free cut score, another error-prone test score, or a pretest score of the same person, one needs the standard error of measurement, not the test-score reliability.

Validity is a conceptually more complex topic than reliability and has been subject to debate until the present day. Validity has two faces. First, as a measurement instrument a test must measure one attribute and not a mixture of different attributes. This goal is best attained when a theory about the attribute drives the test construction process. Modern psychometric latent variable models can be helpful to study the test's validity. The development of a theory is a conceptual endeavor requiring empirical research to find support for the theory. Hence, psychometric models cannot replace the attribute theory but they can lend a helping hand in developing the theory. Second, once a theory-driven test has been constructed the test's practical usefulness for certain applications can be investigated. The two stages of ascertaining construct validity and predictive validity must be distinguished sharply as they serve totally different purposes. Finally, validity is a property of tests and test scores, not of psychometric models, such as classical test theory and item response theory (Henly, this volume).

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Atkins, D. C., Bedics, J. D., McGlinchey, J. B., & Beauchaine, T. P. (2005). Assessing clinical significance: Does it matter which method we use? *Journal of Consulting and Clinical Psychology*, 73, 982–989.
- Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment*, 82, 60–70.
- Bentler, P. A., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika*, 45, 249–267.
- Borsboom, D., Cramer, A. O. J., Kievit, R. A., Zand Scholten, A., & Franić, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 135–170). Charlotte, NC: Information Age.
- Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. Cambridge: Cambridge University Press.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- De Boeck, P., & Wilson, M. (2004). *Explanatory item response models. A generalized linear and nonlinear approach*. New York, NY: Springer-Verlag.
- De Jong, A. E. E., Bremer, M., Schouten, M., Tuinebreijer, W. E., & Faber, A. W. (2005). Reliability and validity of the pain observation scale for young children and the visual analogue scale in children with burns. *Burns*, 31, 198–204.
- Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* [Test theory. Introduction to the theory of the psychological test and its applications]. Houten, the Netherlands: Springer/Bohn Stafleu Van Loghum.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to psychological test theory]. Bern, Switzerland: Huber.
- Gandek, B., Sinclair, S. J., Kosinski, M., & Ware, J. E., Jr. (2004). Psychometric evaluation of the SF-36® health survey in Medicare Managed Care. *Healthcare Financing Review*, 25, 5–25.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59, 12–19.
- Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika*, 42, 567–578.
- Jansen, B. R. J., & van der Maas, H. L. J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology*, 81, 383–416.
- Kruijen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, 12, 321–344.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Ma, C., Chen, S., You, L., Luo, Z., & Xing, C. (2012). Development and psychometric evaluation of the Treatment Adherence Questionnaire of Patients with hypertension. *Journal of Advanced Nursing*, 68, 1402–1413.
- Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, 1, 293–299.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.
- Michell, J. (1999). *Measurement in psychology. A critical history of a methodological concept*. New York, NY: Cambridge University Press.
- Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32, 1–13.
- Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.
- Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling*, 9, 195–212.
- Riley, W. T., Rothrock, N., Bruce, B., Christodolou, C., Cook, K., Hahn, E. A., & Cella, D. (2010). Patient-reported outcome measurement information system (PROMIS) domain names and definitions revisions: Further evaluation of content validity in IRT-derived item banks. *Quality of Life Research*, 19, 1311–1321.

- Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74, 107–120.
- Sijtsma, K. (2009b). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing*, 9, 167–194.
- Sijtsma, K. (2012a). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika*, 77, 4–20.
- Sijtsma, K. (2012b). Psychological measurement between physics and statistics. *Theory & Psychology*, 22, 786–809.
- Sijtsma, K., & Emons, W.H.M. (2011). Advice on total-score reliability issues in psychosomatic measurement. *Journal of Psychosomatic Research*, 70, 565–572.
- Sijtsma, K., Emons, W.H.M., Bouwmeester, S., Nykliček, I., & Roorda, L. D. (2008). Nonparametric IRT analysis of quality-of-life scales and its application to the World Health Organization Quality-of-Life scale (WHOQOL-Bref). *Quality of Life Research*, 17, 275–290.
- Sijtsma, K., & Van der Ark, L. A. (2015). Conceptions of reliability revisited and practical recommendations. *Nursing Research*, 64, 128–136.
- Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. In R. W. Lissitz (Ed.), *The concept of validity. Revisions, new directions, and applications* (pp. 19–37). Charlotte, NC: Information Age.
- Smith, J. A. (1981). The idea of health: A philosophical inquiry. *Advances in Nursing Science*, 3(3), 43–50.
- Smits, D. J. M., & De Boeck, P. (2003). A componential IRT model for guilt. *Multivariate Behavioral Research*, 38, 161–188.
- Ten Berge, J. M. F., Snijders, T. A. B., & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. *Psychometrika*, 46, 201–213.
- Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, 69, 613–625.
- Ten Berge, J. M. F., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika*, 43, 575–579.
- Van der Ark, L. A., Van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement*, 35, 380–392.
- Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.
- Ware, J. E., Jr. (2000). SF-36 health survey update. *Spine*, 25, 3130–3139.
- Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity. Revisions, new directions, and applications* (pp. 65–82). Charlotte, NC: Information Age.

4

CLINIMETRICS

Henrica de Vet and Anna Beurskens

What is clinimetrics?

Clinimetrics is a methodological discipline focused on development and evaluation of instruments used to assess disease and health status. The term “clinimetrics” is indissolubly connected with Alvan Feinstein; he defined clinimetrics as “measurement of clinical phenomena” (Feinstein, 1987). We broadened the definition to “a methodological discipline involved in development and evaluation of clinical measurement instruments” (de Vet, Terwee, Mokkink, & Knol, 2011). Clinimetrics concerns all measurements in the health sciences, including nursing. Examples from the daily practice of nursing and nursing research are body temperature (measured with a thermometer), pain in newborns (measured by standardized observation), mobility or balance (measured by performance tests), and aspects of quality of life (measured by patient self-report).

Clinimetrics overlaps with biometrics and psychometrics, but is distinct in its focus and areas of application. “Biometrics” refers to identification of humans by physical (DNA, fingerprints, iris) and behavioral (gait cadence, voice pitch) characteristics, whereas clinimetrics is more focused on characteristics that are expressions of pathological processes, such as angina pectoris or depression, or health-related behavior in the human body. “Psychometrics” refers to theory and methods used to measure psychological entities such as cognition, emotions, and observable behavior. Biometrics and clinimetrics both rely, to some extent, on the measurement theories of psychometrics (classical test theory and item response theory), especially when multi-item questionnaires are used.

The field of measurement has its origin in different disciplines using various terminologies. Consequently, different terms are used for similar measurement properties and the same term may have different meanings in different fields. The COnsensus-based Standards for selection of health Measurement INstruments (COSMIN; www.cosmin.nl) project reached consensus among participating psychometricians, statisticians, epidemiologists, and clinicians on a taxonomy, terminology, and definitions for many terms as used in health measurements (Mokkink et al., 2010a). We follow COSMIN terminology throughout this chapter.

The central question of this chapter is, “How to assess and select the best instrument for a specific purpose,” from a clinimetric perspective – when many methods are available. The first task is to determine what exactly is to be measured. Next, the type of measurement approaches and practical considerations play a role. Third, the measurement properties of scores obtained

from selected instruments should be appraised. The information on these three steps is integrated in systematic reviews of measurement instruments. We explain where these reviews can be found and how they can be performed.

Defining what to measure

The first step is to define the relevant domains or constructs one intends to measure. The International Classification of Functioning, Disability and Health (ICF) provides a useful framework and standardized language for describing health and health-related states across the domains of body functions and structure, activity limitations, participation restrictions, environmental factors, and personal factors (World Health Organization, 2001). Disturbances of body functions and structure, type and severity of symptoms, physical or mental functioning, and quality of life may all be relevant to evaluating treatment effectiveness. Disease impact can be assessed on all aspects of life or restricted to health-related quality of life. Each focus requires different instruments.

Next, identification of purpose is critical to selection of an appropriate instrument. We distinguish instruments used for diagnosing a disease or level of functional health status from those used for predicting outcome of a specific disease or evaluating treatment effects. For example, the APGAR score is used for assessment of health status in newborns to support clinical decision making in the minutes following birth. However, if intervention is needed, effectiveness isn't assessed by the APGAR score but with instruments targeted at measuring the specific dysfunction of interest.

Measurement approaches

Approaches to measure health-related phenomena are diverse, including imaging techniques, laboratory tests, clinical observations, performance tests, and questionnaires or interviews. For some measurement domains or constructs, there is hardly any choice among instruments to use. For measuring height, a rod or cord is used, and for measuring body weight, a spring or balance scale is used. To assess pain intensity, we just ask patients.

Sometimes instrument choice depends on the exact definition for the domain or construct of interest. An example involves mobility after knee arthroplasty. If range of motion (a body function) is of interest, use of a goniometer is appropriate; when participation in activities that could be performed before the osteoarthritis complaints started is of interest, a questionnaire is selected. An observation or performance test could be used to determine whether the patient can do specific activities. Each approach measures slightly different things: A questionnaire gives answers about what patients think they can do (perceived ability), whereas a performance test shows what a patient can do under standardized conditions. What a person generally does in daily life can be assessed by accelerometers. So, there is a clear interplay between choice of approach and what we exactly want to measure.

Patient-reported outcomes

Patient-reported outcomes (PROs) are “any reports coming directly from patients about how they function or feel in relation to a health condition and its therapy, without interpretation of the patient's responses by a clinician or anyone else” (Patrick et al., 2007). Interviews and questionnaires are used to obtain PROs. It is important to pay attention to the content of the questions and the response options used to measure PROs. It is tempting to think of questionnaires as a quick and easy way to gather information from patients: Just ask them what you want to know.

But this is only true when questions are relatively straightforward. The question, “How many children do you have?” is already more difficult. Should women count all pregnancies, and also count the children who have already died? Even more difficult are questions about feelings, about things that cannot be observed, like depression and anxiety. Called latent variables (constructs), they are usually assessed by multi-item questionnaires. All items aim to measure the same construct, and responses to the item set provide scores that are more reliable than asking a patient “are you depressed?”

Response options and scales of measurement

Clinimetric measurement approaches produce scores in a variety of ways. Laboratory assessments or assessment of walking distance give a direct score on a scale with known units (e.g., mg/dL of hemoglobin in the blood). In the Jebsen–Taylor Hand Function Test, time to complete seven tasks is recorded and can be compared with age-matched norms (Jebsen, Taylor, Trieschmann, & Howard, 1969). MRI, PET scans, cytological or histological slides, electrocardiogram tracings, ultrasound and photographs create images with characteristics assessed using preset criteria that result in scores (measurements) with a variety of characteristics.

Many performance and observation tests and questionnaires make use of prestructured response options that should correspond with the questions posed. Sometimes the meaning of the question is solely determined by the nature of the response options. For example, “Do you have pain?” refers to severity when the response options are *none, a little, a lot* and refers to frequency when response options are *never, sometimes, always*. Response options may create scores on scales with nominal, ordinal, interval, and ratio properties (can be remembered using the acronym “NOIR”).

Nominal measurement uses mutually exclusive classes for response options, for example, blood group O, A, B, or AB. An example of questions with a response option “yes” or “no” include: “Do either of your parents have diabetes?” Nominal responses with two options are “dichotomous.”

Response options may be graded (ordinal; Likert-type format), as when patients are asked to give their opinion about specific statements. For the statement “I have a healthy diet,” patients choose from *strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree*. Likert-type responses usually have three, five, or seven options, with a neutral category in the middle; an even number of options forces a positive or negative response. Response to “Do you have pain?” with response options *none, a little, a lot* is an ordinal scale.

Scores on scales with quasi- or equal-intervals result from many measurement operations (e.g., body temperature, plasma glucose, blood pressure). Differences between systolic blood pressures of 140 and 150 mmHG and between 150 and 160 mmHG are mathematically equal, although the health consequences of the difference may differ. Numeric rating scales and visual analogue scales ask for a single numeric score. Patients can grade pain intensity using an integer from 0 = *no pain at all* and 10 = *unbearable pain*. Or, a 100 mm line is drawn, and patients indicate with an arrow how much pain is being experienced. Now pain is expressed as numbers and is considered to be measured on an interval scale. Scores may be on quasi-interval scales. Total scores from multi-item questionnaires are often treated as quasi-interval or interval measurement, but the scale of the scores is arbitrary when a construct is measured with, say 10 items, and each item has a five-point response option so that the range of total scores is 40.

Measurements with ratio-level characteristics are similar to the interval level. In addition to equal intervals, they have an absolute (true) zero-point. Examples are tumour size and age.

Both the nominal and the ordinal levels use classifications and are categorical variables. Interval and ratio levels scores are continuous outcomes. Whether scores have nominal, ordinal, interval,

or ratio-level characteristics has consequences for statistical analyses (Cliff, 1993; Skrondal & Rabe-Hesketh, 2004; Stevens, 1946). In this chapter, we focus only on the consequences for assessment of measurement properties of scores obtained from instruments.

Practical considerations

Practical considerations are important when choosing an instrument. The issues involve availability and acceptability or feasibility with respect to time, costs, and burden for patient or assessor.

Availability of instruments

There might be high-tech instruments to measure body temperature, but nurses working in the community just do not have access to these. Physical functioning can be assessed by a battery of performance tests, requiring sophisticated apparatus and a gym, but only rehabilitation centers have the facilities and personnel to perform these tests. And often other easier and sufficient alternatives, for example, accelerometers, are available.

Time

In general, in clinical practice, there is little time to perform measurements for research purposes. Therefore, one might easily opt for short tests or questionnaires when conducting research in clinical settings.

Burden

Apart from time, the main burden issues for assessors involve money, as well as expertise, experience, and training. Burden for the patient is also an important consideration. Questions as “how invasive is the diagnostic test?” are actual in the case of colonoscopy, for example. Performance tests may burden the patient in terms of pain or fatigue. Questionnaires may burden patients when they are too long or when they contain emotional and personal questions, or confront patients with impairments, severe limitations, or poor health status. For that reason, research ethics committees usually want to see questionnaires that are used in research.

Capacities of patients need to be considered. Patients who suffer from fatigue or loss of concentration can't fill out long questionnaires. Patients with cognitive problems can't fill out questionnaires that rely on recall. Patients who are depressed may not be able to rightly judge their own health status. Not everybody is able to read, write, or hear. In that case, structured interviews may be a better choice than self-report questionnaires. Language should be tailored to level of cognitive development (children) or educational level (adults). Response options may need to be adapted to the target population by using pictures rather than words (Figure 4.1). In short, we should always judge whether the measurement method is suitable for all subjects in the target population.

Measurement properties

Figure 4.2 shows the taxonomy of properties of clinimetric measurements proposed by the international COSMIN panel (Mokkink et al., 2010a). Domains included are validity, reliability, responsiveness, and interpretability.

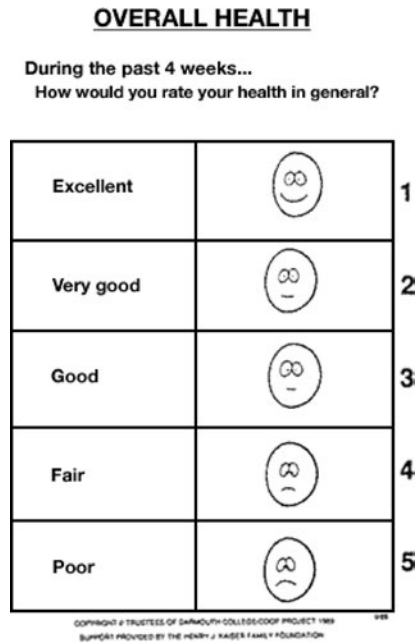


Figure 4.1 An ordinal scale using numbers, words, and pictures to grade overall health. From the Dartmouth-Northern New England Primary Care Cooperative Information Project (www.dartmouthcoopproject.org/coopcharts.html). Used with permission.

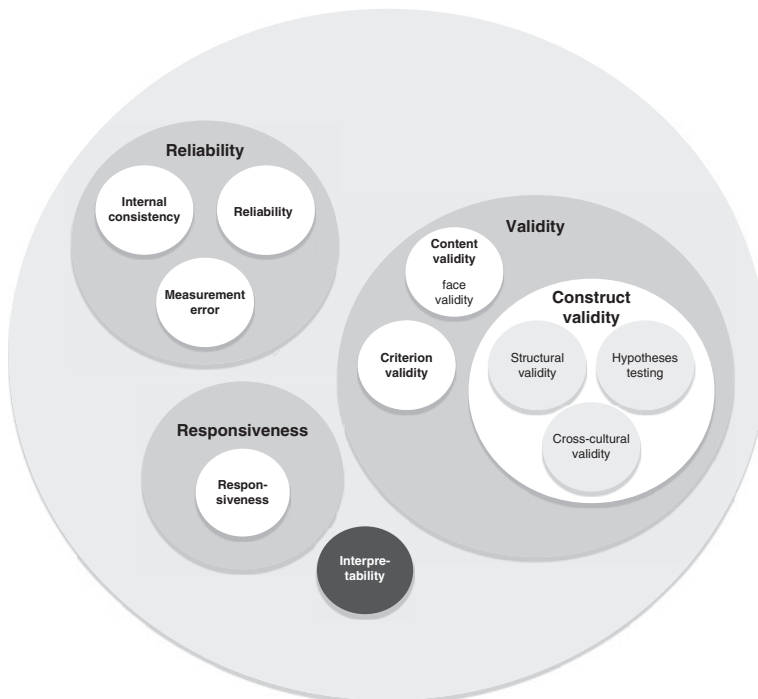


Figure 4.2 The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) taxonomy of relationships among measurement properties (www.cosmin.nl). Used with permission.

Validity

Validity is “the degree to which an instrument measures what it intends to measure.” Face, content, criterion-related and construct validation approaches are all important in clinimetrics.

Face and content validity

Face validity addresses the question, “Does this instrument look like it measures what it is intended to measure?” Questionnaire items should appear to fit the domain or construct. Performance tests may be judged on their fit for a specific population. One might think of shuttle walk test instead of a shuttle run test to assess fitness in an elderly population. An exception of the rule that the face validity should be high is when the assessors purposefully want to “hide” the purpose of the measurement; e.g., in order to avoid socially desirable answers in a questionnaire about behaviour.

Content validation involves assessment of the degree to which items sample a performance domain or construct (Crocker & Algina, 1986), and is usually based on expert opinion about the relevance of items to all aspects of the construct; physical, mental, and motivational aspects of fatigue should be represented in a fatigue item set, for example. Assessment of content validity involves judgement about whether all aspects are covered by the item set, and is used to make decisions such as removing nonrelevant items, improving the wording of the items, or adding new items. Content validation can incorporate expert opinion and input from patients (e.g., Albers et al., 2011; Jiang et al., 2014; Patrick et al., 2011a, 2011b).

Criterion-related validity

Criterion validity involves correspondence between scores obtained from an instrument and a gold standard (scores obtained from a method that perfectly reflects the construct under study). For many constructs, gold standards (or reference standards) do not exist. Examples are symptoms like pain, fatigue, physical functioning, and quality of life.

When gold standards do exist, the question is why another instrument is needed. Reasons are that the gold standard is expensive or cumbersome. For example, consider measurement of proportion fat in total body composition using gold standards like dual energy X-ray absorptiometry or underwater weighing, compared to use of skin fold measurements as a reasonable estimation of proportion body fat. So criterion-related validity is at stake when a gold standard exists, and an alternate instrument has advantages of being faster, cheaper, or less cumbersome.

Construct validity

Construct validation involves gaining evidence in support of inferences from scores obtained from instruments when no gold standard is available. In the COSMIN taxonomy, construct validity is subdivided into hypotheses testing, structural validity, and cross-cultural validity.

Hypothesis testing provides critical evidence about construct validity. Using COSMIN terminology, knowing what the instrument is aimed to measure, hypotheses are formulated about correlations with scores on instruments that measure similar constructs (convergent validity), differences in scores between groups that are known or expected to differ (discriminative validity), or groups where no differences are expected or where no correlations are expected (discriminant analysis). For example, Kluivers et al. (2008) assessed recovery following hysterectomy (vaginal, abdominal, and laparoscopic) using a general quality of life instrument and two quality of recovery instruments at one and four weeks post-op. Based on past findings, they formulated

and tested hypotheses about group differences and relationships among scores obtained using the three instruments to come to a conclusion about which instrument(s) provided the best data.

Structural validation addresses dimensionality of responses to questions on multi-item instruments, considering the theory of the construct. Factor analysis is typically used for structural validation. Structural validation also incorporates assessment of whether a factor analysis model for scores from an instrument is the same in some respects in multiple populations (Jöreskog, 1971). Dimensional, configural, metric, strong, and strict factorial invariance can be tested and are essential for making meaningful comparisons across groups (Gregorich, 2006). An example is whether the Center of Epidemiological Studies Depression Scale (CES-D), developed in the general population, is structurally similar in patients suffering from a specific disorder (e.g., systemic sclerosis; Thombs et al., 2008).

Cross-cultural validation refers to assessment of validity when an instrument is used in another country or culture. First, a process of translation, synthesis, back translation, expert committee review, and pretesting is used (Beaton, Bombardier, Guillemin, & Ferraz, 2000). Then, cross-cultural validation using factor analysis (Gregorich, 2006) or item response theory (e.g., Oude Voshaar et al., 2014) can be used to assess whether items function in similar ways in the cross-cultural groups.

Reliability

Reliability refers to the extent to which scores for patients who have not changed are the same when measurement is repeated under different “conditions” (e.g., using multiple items on a questionnaire, over time, by different persons on the same occasion, by the same persons on different occasions. Sijtsma (this volume) discussed technical issues in estimation of reliability using the classical test theory model. We show the link between reliability and standard error of measurement in clinical measurements, and then focus on test-retest reliability and interrater reliability.

Reliability and standard error of measurement

Reliability and measurement error are related, but different concepts. Standard error of measurement is an absolute indicator of precision that describes variation in observed scores around true scores (in a classical test theory sense). Reliability is a relative measure that describes proportion of true score variance in the total variance of observed scores of a group; thus, reliability is a function of variability in the observed scores in the sample. Reliability can be expressed as:

$$\rho_{TX}^2 = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}, \quad (4.1)$$

where ρ_{TX}^2 is the correlation between true score T and observed score X on a test of interest (one definition of reliability from classical test theory; Lord & Novick, 1968, p. 198), σ_T^2 is the variance of true scores in the population, and σ_E^2 is the error variance in observed scores. The standard error of measurement (SEM) is the square root of σ_E^2 . When the standard deviation of observed scores and correlation between true score and observed score apply to the same group, the SEM, indicated by σ_E , can be obtained with some algebraic manipulation from Equation 4.1 as:

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{TX}}. \quad (4.2)$$

Reliability expresses how well patients can be distinguished from each other on the basis of their test scores despite measurement error. It is easier to distinguish patients from each other in a more heterogeneous sample than in a more homogeneous sample and, therefore, reliability estimates for the same test will vary as a function of the standard deviation of the observed scores in different groups (Lord & Novick, 1968, p. 1999). Thus, reliability concerns distinguishing among patients whereas SEM is relevant to precision of scores.

An example

As part of a national pressure ulcer (PU) prevalence survey in the Netherlands, Braden scale scores for pressure ulcer risk and grade were obtained for homecare clients (Kottner, Halfens, & Dassen, 2009). In the 2008 survey, two scores from separate nurse assessments were available for 339 clients.

Total risk scores reflect six factors contributing to PU development (sensory perception; moisture; activity; mobility; nutrition; friction and shear) rated on scales of 1–4 (except friction and shear, which is rated 1–3). Total scores range from 6–23, with higher scores indicating less risk. Five pressure ulcer grades (ordered categories) were used (from 0 = *no PU* to 4 = *full thickness skin or tissue loss*). For the risk score, differences in pairs of nurse ratings were visualized using Bland-Altman plots (Bland & Altman, 1999), and intraclass correlation (ICC) was used to estimate reliability. For the PU grading, absolute agreement in ratings and agreement corrected for chance were computed.

As shown in Figure 4.3, the Bland-Altman plot shows differences between pairs of nurse ratings for the same homecare client as a function of mean value of the assessments. Nurse ratings (total scores) were the same in 63% of the cases. The size of the circles represents the frequency of the score differences. Assuming the differences between ratings are normally distributed, limits of agreement can be obtained using

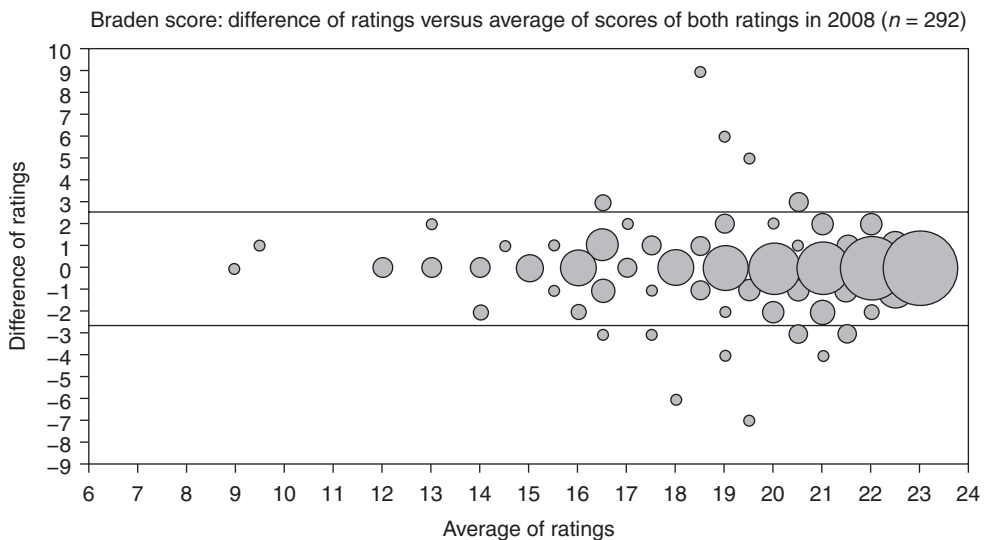


Figure 4.3. A Bland-Altman plot. This Bland-Altman plot shows the difference between two ratings plotted against the average of the ratings. The sizes of the circles represent the frequencies of the difference scores. Nursing ratings were the same in 63% of the cases. The limits of agreement lay at -2.7 and $+2.7$ points on the Braden Scale total score for pressure ulcer risk, in this study. From Kottner et al. (2009). Used with permission.

$$\text{Upper and lower limits of agreement} = \text{Mean Difference} \pm 1.96 \times SD_{\text{difference}}.$$

In this example, there were no systematic differences between the first and second assessment (mean difference was 0), and limits of agreement lay at -2.7 and $+2.7$ points on the Braden total score for PU risk. Interrater reliability estimated using intraclass correlations (ICC) from a one-way random effects model (assuming raters randomly selected from a population of raters) was 0.88 (95% CI $[0.85, 0.91]$). Selection of an appropriate model for estimating the ICC depends on assumptions about raters of interest and whether inferences are made about individual or average ratings (de Vet et al., 2011; Harrison, this volume; McGraw & Wong, 1996).

Note that the magnitude of measurement error can only be assessed for continuous measures. For categorical variables, patients are classified, and it is only possible to determine whether they occur in the same category or not. The term “agreement” is used for this correspondence, sometimes adjusted for chance agreement resulting in Cohen’s kappa. Furthermore, total agreement or specific agreement (i.e., agreement on positive scores and agreement on negative scores, separately) may be estimated (de Vet et al. 2013).

Responsiveness

Responsiveness refers to detection of change over time, clinically important change, or true change in the construct to be measured and many definitions have been proposed (Terwee, Dekker, Wiersinga, Prummel, & Bossuyt, 2003, p. 351). Because responsiveness concerns validity of change scores, it is sometimes called longitudinal validity (Hays & Hadorn, 1992). Assessment of responsiveness is similar to other validation strategies, but focuses on evaluation of hypotheses about change in key health status variables over time or following treatment with known or demonstrated efficacy (Revicki, Hays, Cella, & Sloan, 2008). For a nice example of a responsiveness study, see Kluivers et al. (2008), who compared various instruments to assess recovery after hysterectomy.

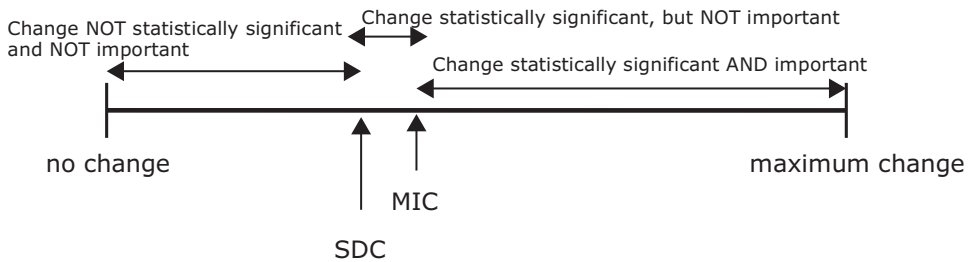
Effect sizes (i.e., mean change divided by standard deviation at baseline) or standardized response mean (i.e., average change score divided by standard deviation of change scores) provide a standardized estimate of magnitude of change when used in conjunction with Cohen’s (1960) criteria, but are less useful as an index of responsiveness. Suppose that a real effect is 1.8 and scores on an instrument only pick up an effect size of 0.8 ; the responsiveness is moderate. Thus, effect size is only useful as an index of responsiveness in a comparative manner with respect to a hypothesized effect, or comparing effect sizes obtained using scores from different instruments in the same study.

Minimally important change or difference (MIC; MID) (referred to as clinically meaningful change in the behavioral sciences literature; Ferrer & Pardo, 2014) and smallest detectable change (SDC) are distinct concepts related to change across two occasions (Terwee, Roorda, Knol, De Boer, & de Vet, 2009). MIC refers to a difference that patients or clinicians consider important. SDC refers to change beyond measurement error. Scores on two occasions from stable patients (e.g., those expected not to change because they received no treatment) can be used to estimate the standard deviation in differences in observed scores when true scores do not change (i.e., all differences are due to measurement error). Guyatt’s responsiveness ratio (GRR; Guyatt, Walter, & Norman, 1987) links minimally important change and smallest detectable change via the standard deviation of the differences among the stable, such that

$$GRR = MIC / SD_{\text{difference}}. \quad (4.3)$$

Assuming a normal distribution for errors of measurement, scores from instruments with $GRR > 1.96$ are considered responsive. Then note that $\pm 1.96 \times SD$ of the difference score are upper

A. Interpretation of change when the MIC is larger than the SDC



B. Interpretation of change when the MIC is smaller than the SDC

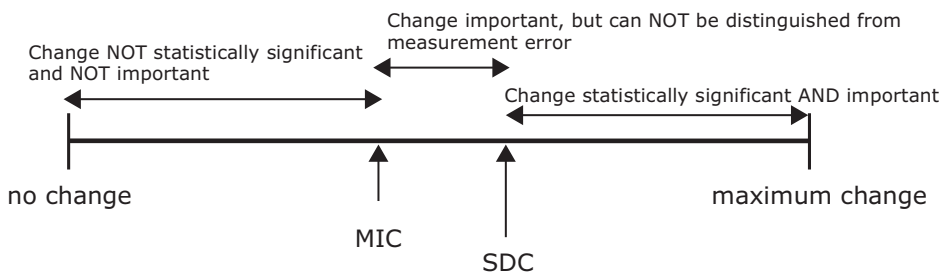


Figure 4.4. Relationships between minimally important change (MIC) and smallest detectable change (SDC). In situation A the smallest detectable change is smaller than the MIC value, and thus MIC can be detected beyond measurement error. In situation B, the MIC value falls within the limits of agreement. A change in between MIC and SDC means that the change would be considered minimally important but we cannot exclude that it is still measurement error, and the measurement instrument should be improved to detect the MIC beyond measurement error. From Terwee et al. (2009). Used with permission.

and lower limits agreement of a Bland and Altman plot (see Figure 4.3). In other words, when observed change lies outside the limits of agreement, it is improbable that observed change is due to measurement error. The absolute value of the upper and lower limits is said to be the SDC, as shown in Figure 4.4. SDC and MIC are both important because they reveal whether scores from an instrument are sufficiently reliable to detect clinical relevant changes. See de Vet et al. (2011) for an overview of methods to assess MIC, which are an active area of methodological development.

Systematic reviews of measurement instruments

Systematic reviews of measurement instruments make inventories of available instruments in a certain field, comparing content, and assessing quality by critically appraising measurement properties in various populations. To assess quality of a systematic review, complete information about

Table 4.1 Instrument databases

Source/site	Content
Center for Outcomes Research and Education www.researchcore.org/gipro/	Gastrointestinal patient-reported outcomes
Educational Testing Service www.ets.org/test_link/about	The Test Collection: database of 25,000 tests
Medical Outcomes Trust www.outcomes-trust.org/instruments.htm	Generic and condition-specific QoL
National Cancer Institute, NIH http://appliedresearch.cancer.gov/paq/	Walking, bicycling
PROMIS www.nihpromis.org	Physical, mental, and social well-being
Rehabilitation Institute of Chicago www.rehabmeasures.org	Patient outcomes across all stages of rehabilitation

Note: NIH = U.S. National Institutes of Health; PROMIS = Patient-Reported Outcomes Measurement Information System; QOL = quality of life.

the search strategy, methods used to perform the review, and use of standards and criteria to assess methodological quality of the instruments included in the review should all be available (www.cosmin.nl; select Systematic Reviews of Measurement Properties tab). This section describes where reviews of measurement instruments can be found and how you can perform such a review yourself.

Finding instruments

Instrument overviews are available in databases, which sometimes list measurement properties obtained from specific populations. Table 4.1 lists database topics, content, and sites.

Performing a systematic review of measurement instruments

A 10-step process similar to that used for intervention effectiveness studies can be used (de Vet et al. 2011).

- 1 **Aim:** To find all evidence on the measurement properties of scores from one or more instruments, evaluate the evidence, and judge qualities of scores obtained from each instrument. If information in the review will be used to select the best instrument for a defined purpose, all instruments and all measurement properties should be included.
- 2 **Search:** To locate all relevant articles, multiple databases should be searched (e.g., MEDLINE, EMBASE, CINAHL). Search terms should include all synonyms for the construct of interest and the target population using a sensitive methodological search filter (Terwee, Jansma et al. 2009).
- 3 **Select:** With regard to papers for review, read abstracts or full articles of the retrieved references, and select by applying strictly formulated inclusion and exclusion criteria. At least two independent reviewers should assess the identified papers.
- 4 **Appraise:** To determine methodological quality of studies, the COSMIN checklist can be used (Mokkink et al. 2010b). This checklist contains items about methodological quality of

measurement studies (comparable to “risk of bias” checklists used in reviews of randomized clinical trials RCTs).

- 5 **Extract data:** To summarize data, systematically tally instrument characteristics, populations studied, and measurement properties of scores. Generalizability of results and judgement about (dis)similarities of populations included in the review can then be made.
- 6 **Define content:** To compare multi-item instruments designed to measure the same or similar constructs, identify topics or elements addressed in questions.
- 7 **Synthesize:** To combine results across studies, identify homogeneous populations, and then statistically pool data or use a best evidence qualitative approach; methods are under development.
- 8 **Conclude:** To draw an overall conclusion about quality of scores from an instrument, or to select the *best* instrument for a particular situation, consider the number of measurement studies, methodological quality of these studies, and consistency of the results.
- 9 **Consider:** To support conclusions, homogeneity is important (similarities in construct measured, purpose, and population). Insufficient data often provide guidance for further research.
- 10 **Report:** To disseminate findings, systematic measurement reviews should include information obtained at each of the prior steps. Extensive tables can be placed in appendices or on websites.

Future developments

In the end, we want to make a note on future developments. Item response theory (IRT; Baldwin & Wainer, this volume) and advances in computer technology enable improved health status measurement through development and maintenance of item banks for measuring specified symptoms and health status domains. The Patient-Reported Outcomes Measurement Information Systems (PROMIS) initiative (www.nihpromis.org) developed item banks (large numbers of calibrated questions) for domains, including pain, fatigue, anxiety, depression, social functioning, physical functioning, and quality of sleep. Computerized adaptive testing (CAT) tools create tests tailored to the individual that produce very reliable scores using only a small number of items – which make their use feasible in clinical practice (given that computers are available).

Conclusion

Choosing the best instrument for a specific purpose starts with defining what exactly needs to be measured. Next, the approach to measurement and practical considerations play a role. Finally, properties of scores obtained using the instrument in a particular population are critical. Systematic measurement reviews provide essential information for making the final choice.

References

- Albers, G., Pasman, H.R.W., Rurup, M.L., de Vet, H.C.W., & Onwuteaka-Philipsen, B.D. (2011). Analysis of the construct of dignity and content validity of the patient dignity inventory. *Health and Quality of Life Outcomes*, 9, 45.
- Beaton, D.E., Bombardier, C., Guillemin, F., & Ferraz, M.B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, 25, 3186–3191.
- Bland, J.M., & Altman, D.G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8, 135–160.
- Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114, 494–509.

- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, and Winston.
- de Vet, H.C.W., Mokkink, L. B., Terwee, C. B., Hoekstra, O. S. & Knol, D. L. (2013). Clinicians are right not to like Cohen's κ . *British Medical Journal*, 346.
- de Vet, H.C.W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. Cambridge: Cambridge University Press.
- Feinstein, A. R. (1987). *Clinimetrics*. New Haven, CT: Yale University Press.
- Ferrer, R., & Pardo, A. (2014). Clinically meaningful change: False positives in the estimation of individual change. *Psychological Assessment*, 26, 370–383.
- Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, 44, S78–S94.
- Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases*, 40, 171–178.
- Hays, R. D., & Hadorn, D. (1992). Responsiveness to change: An aspect of validity, not a separate dimension. *Quality of Life Research*, 1, 73–75.
- Jebsen, R. H., Taylor, N., Trieschmann, R. B., & Howard, L. A. (1969). An objective and standardized test of hand function. *Archives of Physical Medicine and Rehabilitation*, 50, 311–319.
- Jiang, Y., Boyle, D. K., Bott, M. J., Wick, J. A., Yu, Q., & Gajewski, B. J. (2014). Expediting clinical and translational research via Bayesian instrument development. *Applied Psychological Measurement*, 38, 296–310.
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409–426.
- Kluijvers, K. B., Hendriks, J. C., Mol, B. W. J., Bongers, M. Y., Vierhout, M. E., Brölmann, H. A. M., & de Vet, H. C. W. (2008). Clinimetric properties of 3 instruments measuring postoperative recovery in a gynecologic surgical population. *Surgery*, 144, 12–21.
- Kottner, J., Halfens, R., & Dassen, T. (2009). An interrater reliability study of the assessment of pressure ulcer risk using the Braden scale and the classification of pressure ulcers in a home care setting. *International Journal of Nursing Studies*, 46, 1307–1312.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- McGraw, K. O., & Wong, S. P. (1996) Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–46.
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., . . . de Vet, H. C. W. (2010a). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, 63, 737–745.
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., . . . de Vet, H. C. W. (2010b). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research*, 19, 539–549.
- Oude Voshaar, M. A. H., ten Klooster, P. M., Glas, C. A. W., Vonkeman, H. E., Taal, E., Krishnan, E., . . . van de Laar, M. A. F. J. (2014). Calibration of the PROMIS Physical Function Item Bank in Dutch patients with rheumatoid arthritis. *PLOS ONE*, 9, e92367.
- Patrick, D. L., Burke, L. B., Powers, J. H., Scott, J. A., Rock, E. P. Dawisha, S., . . . Kennedy, D. L. (2007). Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value in Health*, 10, S125–S137.
- Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., & Ring, L. (2011a). Content validity – Establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force Report: Part 1 – Eliciting concepts for a new PRO instrument. *Value in Health*, 14, 967–977.
- Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., & Ring, L. (2011b). Content validity – Establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force Report: Part 2 – Assessing respondent understanding. *Value in Health*, 14, 978–988.
- Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology*, 61, 102–109.

- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling. Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680.
- Terwee, C. B., Dekker, F. W., Wiersinga, W. M., Prummel, M. F., & Bossuyt, P. M. M. (2003). On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Quality of Life Research*, 12, 349–362.
- Terwee, C. B., Jansma, E. P., Riphagen, I. I., & de Vet, H. C. W. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research*, 18, 1115–1123.
- Terwee, C. B., Roorda, L. D., Knol, D. L., De Boer, M. R., & de Vet, H. C. W. (2009). Linking measurement error to minimal important change of patient-reported outcomes. *Journal of Clinical Epidemiology*, 62, 1062–1067.
- Thombs, B. D., Hudson, M., Schieir, O., Taillefer, S. S., Baron, M., & Canadian Scleroderma Research Group. (2008). Reliability and validity of the Center for Epidemiological Studies Depression Scale in patients with systemic sclerosis. *Arthritis Care & Research*, 59, 438–443.
- World Health Organization. (2001). *International classification of functioning, disability and health: ICF short version*. Geneva, Switzerland: Author.

5

ITEM RESPONSE THEORY

A statistical theory of measurement based on fungible items

Peter Baldwin and Howard Wainer

Item response theory (IRT) is a family of mathematical models that provide a stochastic description of what happens when a person meets an item. Items can be part of a test, a survey, or any other questionnaire and the responses can be binary, graded, categorical, continuous, or any combinations of these. There are many variations of IRT models from which the user can select to score a given instrument best for a given purpose.

IRT posits that latent respondent traits (personal attributes or characteristics that are not directly observable) predict responses to items used to elicit trait-relevant information. For example, physical functioning is a latent trait that may be used to predict responses to items in the PROMIS physical function item bank (Fries, Cella, Rose, Krishnan, & Bruce, 2009). IRT is a measurement theory because a respondent's *level* on the trait can be estimated once the probabilistic relationship between trait and response options is quantified.

In this chapter, IRT is presented as an alternative to true score theory in the development and use of measurement instruments. In addition to introducing key concepts and methods in IRT, we describe in some detail four important classes of IRT models, approaches to estimating item and person parameters along with some of the issues and challenges that arise during their estimation, and the evaluation of model-data fit. IRT software is identified.

Historical context

True score theory – sometimes, problematically, called *classical test theory* (Wainer, 2011) – which is described by Sijtsma in this volume, has proven to be a powerful tool for researchers and measurement specialists over the past century. Yet, despite its utility, it has a number of well-known shortcomings (Lord, 1980; Hambleton & Swaminathan, 1985; Embretson & Reise, 2000). Chief among these is that respondent characteristics are only interpretable with respect to a specific test or instrument and that test and item characteristics are only interpretable with respect to a specific sample of respondents. Further, fundamental to true score theory (specifically to its value as a theory of measurement) are the related concepts of reliability and standard errors of measurement, which are somewhat unrealistically defined in terms of strictly parallel tests and most often without conditioning on respondent traits. Finally, true score theory relies on a weak model and thus often requires large datasets to use effectively or, put another way, true score theory lacks the kinds of strong assumptions that can compensate for the absence of large data sets.

Development of item response theory (sometimes called latent trait theory, strong true score theory, or modern test theory) began in earnest in the early 1950s (Lord, 1952, 1953a, 1953b) and can be seen as a response to perceived shortcomings of true score theory as articulated at that time (Gulliksen, 1950). As noted above, IRT comprises a family of mathematical descriptions of what happens when a respondent interacts with an item (e.g., a test question, performance task, survey question, or a therapeutic). Specifically, these models describe the probability of a given response to an item as a function of certain item characteristics and respondent latent traits.¹

In IRT, like many regression models, when the model holds exactly, item and respondent parameters are independent of the sample of respondents and the sample of items, respectively. Parameters with this property are said to be *invariant*. The implication of invariant respondent parameters – like proficiency – is that examiners may present unique non-random sets of items to different respondents and still compare their performances. Likewise, invariant item characteristics are item characteristics that are not dependent on the sample of respondents to which they are administered, which allows items to be compared even when their characteristics are estimated using non-equivalent samples. Note that strictly speaking, *invariance is a property of parameters – not their estimates*. Thus, for example, when proficiency distributions are wildly different, we should not expect their associated respective item parameter *estimates* to closely resemble each other. In this chapter we introduce a number of attractive features of IRT; however, invariance is foundational to many of the topics we discuss.

IRT models

While there exist over 100 unique IRT models, familiarity with all of these models is not necessary to profitably use IRT. Indeed, the vast majority of IRT models remain of principal interest to researchers – not practitioners. Curious readers are directed to books by van der Linden and Hambleton (1997), Nering and Ostini (2010), and Fox (2010), where many of these models are described in detail. Here, our interest is limited to four of the best-known IRT models, which have a long history of productive application to a broad range of measurement problems: the three logistic models for dichotomous (i.e., binary) response data, the 1-PL, 2-PL, and 3-PL, and the graded response model (GRM) for ordered categorical response data (also called *rating scale* data). The former three are the most widely used of all IRT models and were first described to greatest notice by Birnbaum in Lord and Novick's (1968) seminal book *Statistical Theories of Mental Test Scores*. The latter, developed by Samejima (1969), was not only the first but it remains among the most widely – if not *the* most widely – used polytomous IRT model.

Most IRT models posit that a single dominant respondent trait motivates – at least to some extent – all responses.² This is an assertion that the complete latent space is *unidimensional*. Indeed, it is reasonable to view IRT as a formal expression (as it happens, a very useful one) of early notions that test items all ought to somehow measure the same thing (Loevinger, 1947). The assumption of unidimensionality can and should be tested. Details about the many specific methods for testing unidimensionality lie outside the scope of this chapter but common approaches include using linear and non-linear factor analytic methods, testing the closely related and in some sense identical assumption of conditional independence (described below), and examining goodness of fit more generally (also discussed below). Brief non-technical descriptions of these and other approaches can be found in Hambleton, Swaminathan, and Rogers (1991), Swaminathan, Hambleton, and Rogers (2007), and Reeve et al. (2007).

Applications of IRT in nursing research are distinct from the educational testing context in which IRT was largely developed. For this reason, some readers may find the vocabulary used

to describe IRT confusing. For instance, it is customary to describe IRT models as giving the probability of a *correct* response conditioned on a specific level of a respondent's trait; however, *correctness* may not be the best description of responses to questions about health behaviors or patient-reported outcomes. We avoid confusion here by referring to responses as keyed *positively* or *negatively*, but other conventions are more intransigent. For instance, in educational testing (or on the National Council of State Boards of Nursing Licensure Examination, the NCLEX), the respondent trait of interest typically conforms to the popular use of the word *proficiency* (e.g., mathematical skill, reading comprehension, nursing skill). In contrast, within nursing and health *research*, traits such as physical functioning or disability (e.g., Hays et al., 2007), dyspnea (e.g., Yount et al., 2011), death anxiety (e.g., Gómez, Hidalgo, & Tomás-Sábado, 2007) or other patient reported outcomes or health-related quality of life measures are the focus of measurement. While *proficiency* would seem less appropriate in this context, because there is no obvious substitution that better encompasses the breadth of potential respondent traits, we unenthusiastically follow convention and use *proficiency* to denote all such traits. Readers are urged to remain cognizant of the diverse range of traits that proficiency may refer to in this context. Likewise, when describing item difficulty, we found no substitute term that improved clarity. Thus, with some reluctance we follow convention on this point too, but remind readers that in many contexts difficulty merely describes the relative propensity of an item to elicit a negative response.

1-PL model for dichotomous data

In one of its simplest forms, IRT characterizes each item only by its difficulty (usually denoted b) and each respondent by a single trait (usually denoted θ). Using a logistic function,³ the IRT model then gives the probability of a positive response in terms of the difference between difficulty, b , and proficiency, θ , both of which are unobservable. One version of this model was first developed and popularized by the Danish mathematician Georg Rasch (1901–1980), and so is sometimes termed the Rasch Model in his honor. Alternatively, in its logistic form it is referred to as the one-parameter logistic model (or the 1-PL for short) because it characterizes each item with just a single parameter (difficulty = b).⁴ We denote it the 1-PL to reinforce this model's position within the parametric family of logistic IRT models. The 1-PL is given by:

$$P(\theta) = \frac{e^{(\theta-b)}}{1 + e^{(\theta-b)}}, \quad (5.1)$$

where $P(\theta)$ is the probability of someone with proficiency θ responding positively to an item of difficulty b . For example, suppose a respondent is asked to endorse/not endorse a given item related to pain. Here, proficiency could refer to a respondent's level of pain, which we might describe as *pain severity*, whereas item difficulty could describe the level to which an item is *not* painful or (somewhat awkwardly) its *painlessness*. For example, a relatively painless item could be “Because of my pain, I find breathing difficult” whereas a relatively painful item (i.e., low painlessness) could be “Because of my pain, I find exercise difficult.” Here, only respondents with high levels of pain are likely to endorse the first relatively painless item while most respondents – even those experiencing very little pain – are likely to endorse the latter relatively painful item. Equation 5.1 tells us the probability that someone with a pain severity of θ would endorse an item with a painlessness level of b .

The structure of the 1-PL model is depicted in Figure 5.1, which shows the 1-PL for two items of different difficulty. These curves are called *item characteristic curves* (ICCs; also sometimes referred to as *trace lines* or *item response functions*). Note that the ICCs for this model do not

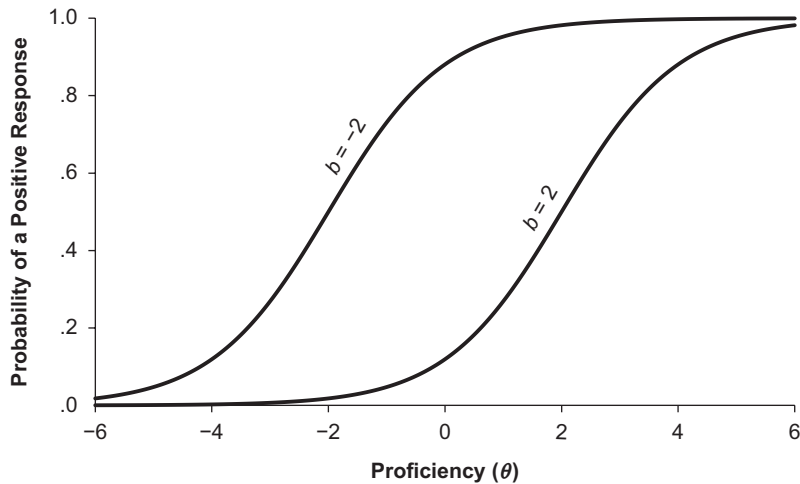


Figure 5.1 Item characteristic curves for two 1-PL items.

intersect at any finite value of θ . This important characteristic of the 1-PL distinguishes it from other more general models. Further, note that when $\theta = b$, the exponential function, $e^{(\theta-b)} = e^0 = 1$ and Equation 5.1 yields the probability $P(\theta) = .5$. Thus, an item difficulty parameter, b , is the point on the proficiency scale at which respondents have a 50% chance of responding positively.

2-PL model for dichotomous data

The 1-PL has many attractive features, and the interested reader is referred to the writings of Georg Rasch (1960) and of Ben Wright (e.g., Wright & Stone, 1979) for convincing descriptions of its efficacy. Nevertheless, it is a strong model with a number of assumptions that can be difficult to meet in practice. Thus, for many applications of IRT to predetermined domains of items, it has been found that more general models fit the observed data much better. A common cause of model-data misfit is that item response data do not always conform to the 1-PL constraint that ICCs never intersect. When this occurs, the measurement specialist must choose between two options. One is to simply delete items that do not fit the 1-PL. This approach is preferred by some practitioners who believe that the benefits of the 1-PL outweigh the disadvantages of excluding misfitting items.⁵ The second less-Procrustean option is to generalize the model to allow ICCs to intersect one another. This is done by adding a second parameter for each item. This parameter, usually denoted a , is proportional to the maximum slope of its item's characteristic curve. The resulting mathematical model, which now contains two parameters per item, is denoted the 2-PL and looks similar to the 1-PL:⁶

$$P(\theta) = \frac{e^{a(\theta-b)}}{1 + e^{a(\theta-b)}} \quad (5.2)$$

Once again our intuition is aided by seeing plots of the ICCs achievable with this more general model. Figure 5.2 shows three 2-PL ICCs for items with the same b parameter ($b = 0$), which demonstrate the variation in slopes sometimes seen in practice (when proficiency is scaled to have unit variance). Shown is an item that has high discrimination ($a = 3$), average discrimination ($a = 1.5$), and lower than average discrimination ($a = .75$). Moving from left to right, note that the slope of each ICC gradually increases until it reaches its maximum and then

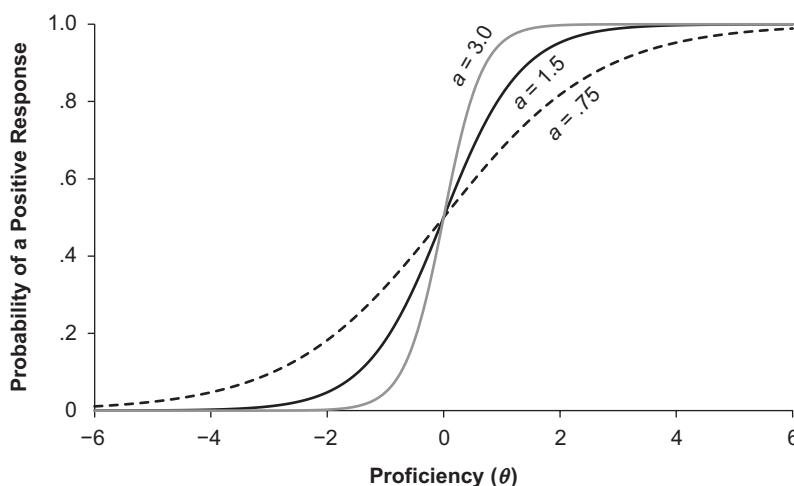


Figure 5.2 Item characteristic curves for three 2-PL items with equal difficulties.

begins to decline. This maximum, the ICC's inflection point, always occurs at an item's difficulty value (in this case at $b = 0$) and for the 1-PL and 2-PL is equal to one-fourth the value of the a -parameter.⁷

The slope of an item is the rate at which the probability of a positive response increases as a function of proficiency – what we might describe as the *steepness* of an ICC. Relatively large slopes correspond to relatively short intervals along the proficiency scale in which $P(\theta)$ moves from nearly zero to nearly one (i.e., in which respondents at one end of such an interval are expected to respond negatively while respondents at the other end are expected to respond positively). The shorter these intervals, the more similar respondents can be with respect to proficiency and still be expected to perform differently. Thus, items with large slopes are said to better *discriminate* among respondents than items with smaller slopes and it is for this reason that a -parameters – being proportional to the maximum slope of an item – are sometimes called *discrimination parameters*. Note, however, that highly discriminating items aren't very useful unless the proficiency interval in which their $P(\theta)$ moves from nearly zero to nearly one corresponds to a region of the proficiency scale that is of interest to the examiner. Indeed, when the region of interest is large, items with more gradual slopes will be more generally useful. Since item difficulty corresponds to an item's maximum slope, items are often selected with difficulties near or within various regions of interest.

3-PL model for dichotomous data

With the addition of the discrimination parameter, the 2-PL greatly expanded the applicability of IRT. Many sets of items that could not fit under the strict equal a -parameter assumption of the 1-PL fit this more general model well. Still, another common source of model-data misfit remains: selected response items – like the ubiquitous multiple-choice question – create the opportunity for respondents to answer positively by guessing. Neither of the two models discussed so far allows for guessing – if a respondent answers positively, it is attributed to their latent proficiency alone. This outcome can be avoided by using a still more general model fully explicated by Allan Birnbaum in Lord and Novick's (1968) comprehensive text. It adds a

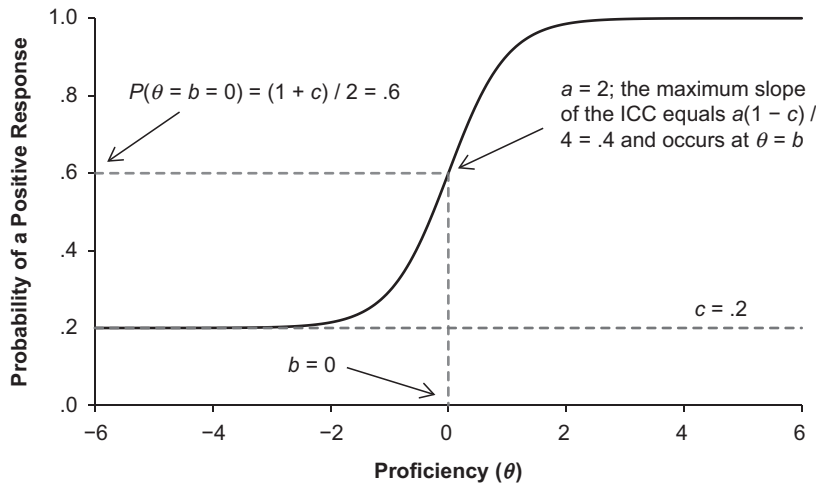


Figure 5.3 Item characteristic curve for a 3-PL item.

third parameter, c , that represents a Bernoulli floor on the probability of responding positively. The resulting model, not surprisingly called the *three parameter logistic model* or simply the 3-PL, is shown in Equation 3.

$$P(\theta) = c + (1 - c) \frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}. \quad (5.3)$$

Once again, we can get a better feel for the structure of the 3-PL once we view a plot of a typical ICC. Such a plot is shown in Figure 5.3:

Note that the ICC in Figure 5.3 has a non-zero lower asymptote. This lower bound reflects the probability of low-proficiency respondents to answer positively. This non-zero probability is principally attributed to item properties that elicit guessing (such as offering only a small number of response options); however, construct-irrelevant respondent traits – notably, one's propensity to guess – are also sometimes acknowledged, although they are typically assumed to be the same across respondents within a given population. For these reasons, the c -parameter is typically denoted the *guessing* or *pseudo-guessing parameter*.

The graded response model for polytomous data

In many instances – perhaps especially in nursing and health science research – response data will be collected that are not dichotomous. For example, it is common on questionnaires to have respondents choose a response from among some number of discrete possibilities. Data like these are called *polytomous* (or sometimes – ungrammatically and typically in older publications – *polychotomous*) and can be modeled using special IRT models designed for this purpose. The Graded Response Model (GRM; Samejima, 1969) is one such model that is widely used to model a particular kind of polytomous data, *ordered* categorical data (sometimes denoted *rating scale* data). Likert-type responses (responses that indicate the *extent* to which a respondent endorses a particular statement – e.g., *strongly disagree*, *disagree*, *neither disagree nor agree*, *agree*, or *strongly agree*) are an example of ordered response categories that will be familiar to most readers.

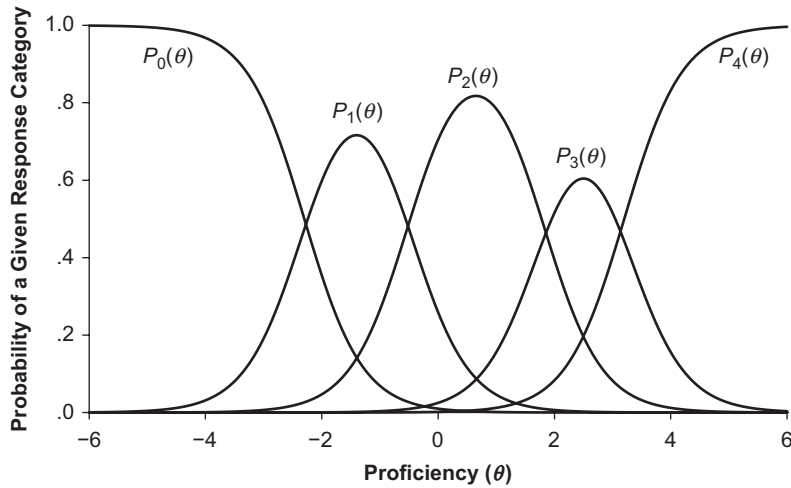


Figure 5.4 Item response category characteristic curves (IRCCCs) for a 5-category item. Each curve shows the probability of a respondent with proficiency θ responding in a given category.

Rather than begin by introducing the relevant equations, some readers may find it easier if we introduce the model by plotting the probability functions for each response category associated with a given item. In contrast to an *item characteristic curve* (ICC), in the context of polytomous IRT models, each response function is referred to as an *item response category characteristic curve* (or simply an IRCCC). Figure 5.4 plots the IRCCCs for an item with 5 categories. Note that categories are ordered and numbered 0 to 4.

Of particular interest in Figure 5.4 is that the form of the IRCCCs is not consistent across score categories: category 0 is monotonic decreasing, category 4 is monotonic increasing, and categories 1, 2, and 3 are unimodal. These variations in shape across IRCCCs complicates the IRT model considerably compared to the dichotomous models introduced above. For this reason, the GRM is reformulated as a set of dichotomous response functions denoted *boundary characteristic functions*. In the 5-category case, six boundary characteristic functions are asserted: four that take the familiar logistic form of the 2-PL with parameters that must be estimated, and 2 that follow logically from the structure of the model and are defined a priori. These six boundary characteristic functions are described in Table 5.1.⁸

Table 5.1 is clarified considerably with the help of a figure. Figure 5.5 plots all six boundary characteristic curves. Note that because the form of the four *estimated* boundary functions, $P_{\geq 1}(\theta)$ through $P_{\geq 4}(\theta)$, share the same form as the 2-PL model, their parameters a and b may be interpreted in the same manner described above; however, for a given item, these parameters are constrained in two ways. First, because the GRM describes *ordered* response categories, it follows that $b_1 < b_2 < b_3 < b_4$. Second, because each boundary function $P_{\geq 1}(\theta)$ through $P_{\geq 4}(\theta)$ (for a given item) shares a common discrimination parameter, a , these four boundary functions are parallel. The boundary characteristic functions for the first and last response options are constants. Note that as indicated by their respective subscripts, all boundary functions – even those that are constants – represent the probability of responding in a given score category *or higher*.

Once the boundary characteristic function parameters are estimated, the IRCCC for category k is given by:

$$P_k(\theta) = P_{\geq k}(\theta) - P_{\geq (k+1)}(\theta). \quad (5.4)$$

Table 5.1 Boundary characteristic functions for 5-category GRM items

Expression	Rationale	Description
$P_{\geq 0}(\theta) \equiv 1$	$P_{\geq 0}(\theta) = P_0(\theta) + P_1(\theta) + \dots + P_4(\theta) = 1$	The probability of responding in category 0 or higher
$P_{\geq 1}(\theta) = \frac{1}{1 + e^{-a(\theta - b_1)}}$	$P_{\geq 1}(\theta) = P_1(\theta) + P_2(\theta) + P_3(\theta) + P_4(\theta)$	The probability of responding in category 1 or higher
$P_{\geq 2}(\theta) = \frac{1}{1 + e^{-a(\theta - b_2)}}$	$P_{\geq 2}(\theta) = P_2(\theta) + P_3(\theta) + P_4(\theta)$	The probability of responding in category 2 or higher
$P_{\geq 3}(\theta) = \frac{1}{1 + e^{-a(\theta - b_3)}}$	$P_{\geq 3}(\theta) = P_3(\theta) + P_4(\theta)$	The probability of responding in category 3 or higher
$P_{\geq 4}(\theta) = \frac{1}{1 + e^{-a(\theta - b_4)}}$	$P_{\geq 4}(\theta) = P_4(\theta)$	The probability of responding in category 4 or higher
$P_{\geq 5}(\theta) \equiv 0$	$P_{\geq 5}(\theta) = P_5(\theta) \equiv 0^*$	The probability of responding in category 5 or higher

Note: Response categories are number 0 to 4.

*The fifth category, category 4, is the highest possible category; thus, a response of 5 or higher is not possible. This boundary characteristic function is defined merely to simplify certain calculations.

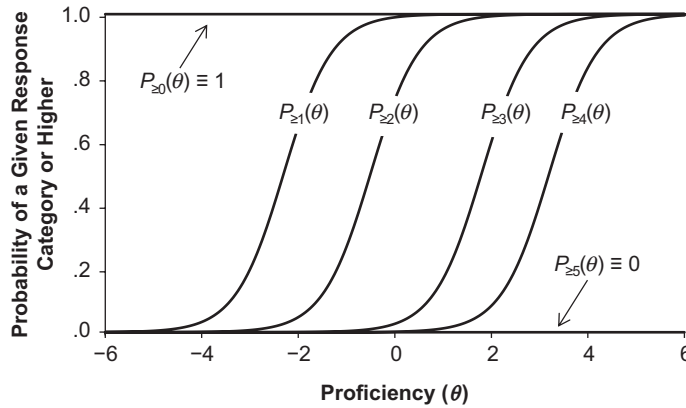


Figure 5.5 Boundary or operating characteristic functions for a 5-category item. Each curve shows the probability of responding in a certain category or higher.

This probability of responding in a particular category k is simply the probability of responding in that category or higher ($P_{\geq k}(\theta)$) minus the probability of scoring in the next category up or higher ($P_{\geq (k+1)}(\theta)$). Further, Samejima (1969) observed that the modes of the unimodal IRCCCs (e.g., $P_1(\theta)$, $P_2(\theta)$, and $P_3(\theta)$ from Figure 5.4) are given by:

$$\max(P_k(\theta)) = \frac{b_k + b_{k+1}}{2}, \quad (5.5)$$

where $\max(P_k(\theta))$ is undefined for $k = 0$ and $k = K$ (where K is the highest response category – e.g., in our 5-category example, $K = 4$). (When $k = 0$ or $k = K$, their respective IRCCC will

be monotonic and no finite maximum exists.) Finally, as implied in Table 5.1 (row 1), note that the probability of responding in *any* of the categories (but none in particular) is, of course, equal to 1:

$$P_{\text{any category}}(\theta) = \sum_{k=0}^K P_k(\theta) = \sum_{k=0}^K [P_{\geq k}(\theta) - P_{\geq (k+1)}] = P_{\geq 0}(\theta) = 1. \quad (5.6)$$

Indeterminacy and the problem of developing a common metric

In Equations 5.1, 5.2, and 5.3, as well as those shown in Table 5.1, we can replace a with $a^* = a / \gamma$, b with $b^* = b\gamma + \eta$, and θ with $\theta^* = \theta\gamma + \eta$, where γ and η are the slope and intercept of some (arbitrary) linear transformation, without affecting the probability of a positive response (i.e., $a^*(\theta^* - b^*) = a(\theta - b)$). This is problematic because there is no basis for preferring one set of parameters over the other – they produce identical estimates of the probabilities of positive responses and provide exactly the same fit to observed data. This scale indeterminacy – called the *identification problem* – should not come as a surprise: because proficiency, difficulty, and discrimination are unobserved, it follows that they *must* be arbitrarily scaled. (This also illustrates that when the same parameters are scaled differently, they can only be considered invariant up to a linear transformation.)

Various conventions and practices exist to identify IRT models but the most common is to scale proficiency such that θ has a mean equal to zero and a standard deviation equal to 1 in some reference group of respondents.⁹ Provided our reference group is thoughtfully chosen, this approach allows us to understand at-a-glance the structure of our results. Regrettably, it also means that unless proficiency distributions are truly equal across samples, independently scaled samples standardized in this way cannot be assumed to share a common scale (because the mean and standard deviation of their proficiency distributions have been made to be identical artifactually). Thus, it is crucial that steps be taken to develop a common metric before comparisons across independently scaled estimates can be made.

The problem of developing a common metric is sometimes referred to as an equating problem; although it is more accurately described as a *scaling problem* (a given parameter expressed on two independent scales is nevertheless the same parameter). When samples are independently scaled, developing a common metric is generally achieved by having a common (or randomly equivalent) set of parameters appear on each scale that can be used to estimate the linear relationship between the scales and thus the linear transformation necessary to put one set of estimates onto the scale of the other. Often, this is accomplished by presenting a common set of items – an *anchor test* – to each sample. Several well-known strategies for estimating this transformation function are described in Hambleton and Swaminathan (1985) and Kolen and Brennan (2004).

Estimation

Calibration describes the estimation of respondent and item parameters for a given IRT model. This is a difficult task because neither the item parameters (a , b , and c) nor the respondents' proficiencies (θ) are known. For this reason, calibration is almost always accomplished using specialized software. Indeed, it is our expectation that readers of this chapter – like most psychometricians – will utilize such software to estimate their model parameters; however, much of the available software requires some understanding of the general problem of estimation to be used judiciously. More generally, it is our experience that users of IRT (both researchers and consumers of research) profit from familiarity with the concepts and vocabulary related to the problem of estimation.

The 3-PL model is widely used for large-scale testing applications and is the most general dichotomous model among those we've introduced; therefore, it is a sensible choice for introducing some of the key concepts underlying estimation and we shall confine the discussion that follows to this model alone. We focus on the method of maximum likelihood (along with some common variations). While this method is *not* the most widely used, it is the basis for the other commonly used methods and as such provides the best conceptual foundation for understanding the general problem of estimation in the context of IRT.

Estimating proficiency and the concept of conditional independence

For pedagogical purposes, let's assume for the moment that we have (somehow) already estimated the three item parameters (a , b , and c) for a given set of items, and subsequently administered this calibrated item set to a sample of respondents. Thus, our task is simply to estimate the proficiency (θ) for each respondent given their responses and the item parameter estimates that are already known to us. (We will return to the problem of joint estimation shortly.)

Let respondent n be some respondent for which we wish to estimate proficiency, θ . To accomplish this, we need to define three new symbols: (i) let $\mathbf{x}_n$ be the vector of scored item responses for respondent n , in which each response is coded 1 if positive, and 0 otherwise; it has elements $\{x_{ni}\}$, where items are indexed by i (e.g., $[x_{n1} \ x_{n2} \ \dots \ x_{nI}] = [1 \ 0 \ \dots \ 0]$, wherein respondent n answers item 1 positively, item 2 negatively, and so on until item I , which is also answered negatively); (ii) let β_i be the item parameter vector (a_i, b_i, c_i) for item i , which is a vector component of the matrix of all item parameters $\mathbf{B}$; and (iii) let $Q(\theta)$ be the probability of responding *negatively*: $Q(\theta) \equiv 1 - P(\theta)$. The conditional probability of the response vector $\mathbf{x}_n$ for a given θ_n and $\mathbf{B}$ (the likelihood) is shown in Equation 5.7:¹⁰

$$L(\mathbf{x}_n | \theta_n, \mathbf{B}) = \prod_i P_i(\theta_n)^{x_{ni}} Q_i(\theta_n)^{1-x_{ni}}. \quad (5.7)$$

Note that when $x_{ni} = 1$, the first multiplicand in Equation 5.7, $[P_i(\theta_n)^{x_{ni}}]$, reflects the ICC for positive responses while the second multiplicand, $[Q_i(\theta_n)^{1-x_{ni}}]$, simply evaluates to 1. In contrast, when $x_{ni} = 0$, it is the first multiplicand in Equation 5.7, $[P_i(\theta_n)^{x_{ni}}]$, that evaluates to 1 while the second, $[Q_i(\theta_n)^{1-x_{ni}}]$ equals the probability of a negative response. Thus, Equation 5.7 is merely the product of the model-generated probabilities for each item response in vector $\mathbf{x}_n$ or, in other words, it is the likelihood – assuming conditional independence – of respondent n 's answers to the I items, $\mathbf{x}_n$, for a given θ_n and $\mathbf{B}$.

Further, note that for a specific response vector $\mathbf{x}_n$, proficiency θ_n , and item parameter matrix $\mathbf{B}$, the likelihood shown in Equation 5.7 yields a single value; however, we may also evaluate Equation 5.7 as a function of θ_n . Figure 5.6 shows this graphically for a two-item test, in which a respondent gives a positive response to the first item and a negative response to the second item. The probabilities for each of these occurrences as a function of θ_n are shown in the top and middle panels of Figure 5.6 respectively.

Thus, if we know $\mathbf{B}$, the item parameters, we can look upon Equation 5.7 for a fixed response pattern $\mathbf{x}_n$, as the likelihood function $L(\theta | \mathbf{x}_n)$ of θ given $\mathbf{x}_n$; its value at any given θ indicates the relative likelihood that $\mathbf{x}_n$ would be observed if θ were the true value. It follows that Equation 5.7 conveys the information about θ contained in the data, and serves as a basis for estimating θ by means of maximum likelihood. The *maximum likelihood estimate* of θ is merely the mode of the likelihood. Stated graphically in terms of Figure 5.6, it is the value of θ associated with the highest point on the likelihood (the bottom panel in Figure 5.6), which in this example occurs near $\theta = -0.1$. As specified in Equation 5.7, this likelihood was obtained by multiplying the curve

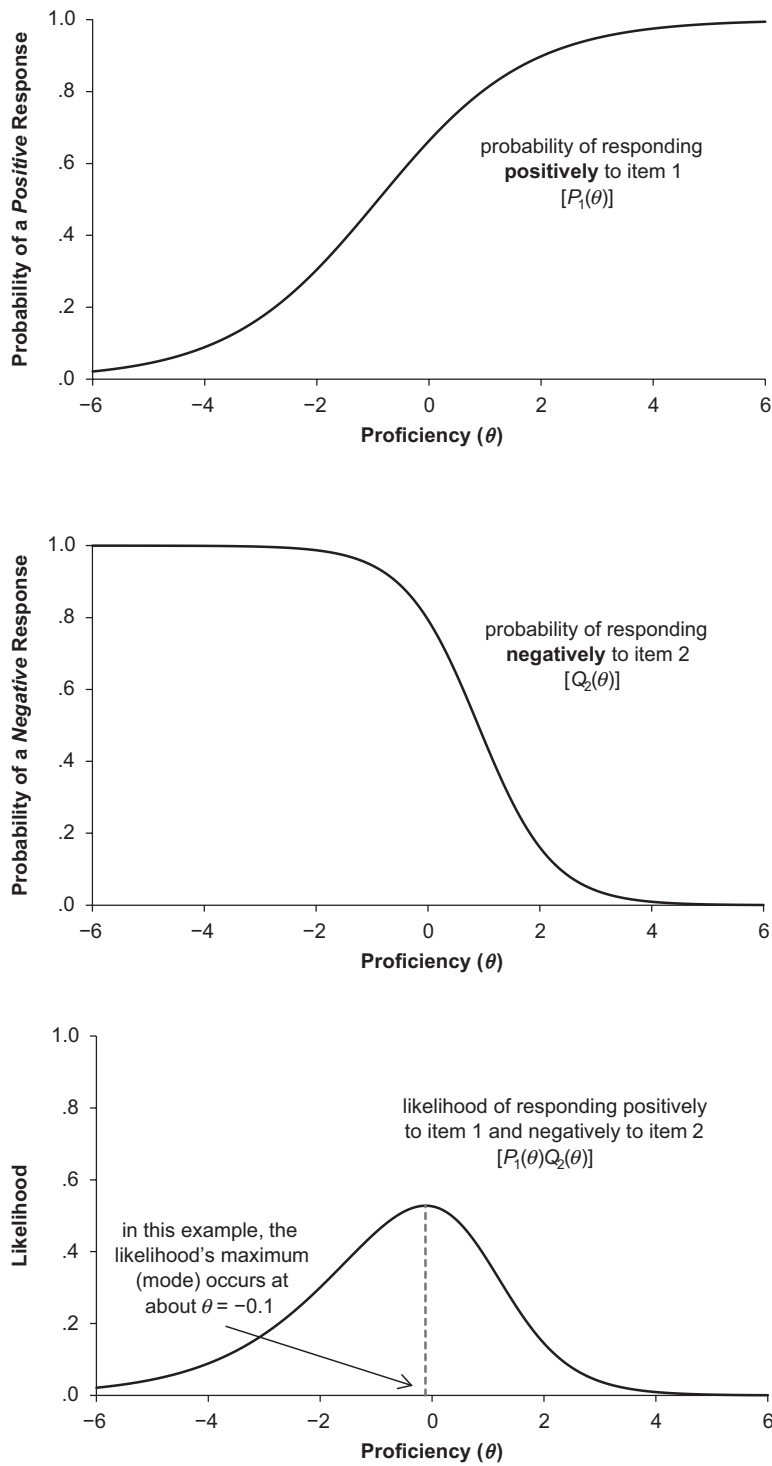


Figure 5.6 The likelihood of responding positively to item 1 and negatively to item 2 as the product of each item's response function.

in the top panel by the curve in the middle panel.¹¹ While finding the value of θ that maximizes Equation 5.7 cannot be done in closed form (i.e., there is no known analytical solution), it is in principle straightforward using an iterative method (most often, Newton–Raphson); however, in practice many problems can arise. These problems are discussed in greater detail below.

On conditional independence

In order for the product shown in Equation 5.7 to describe the likelihood of a particular response vector, item responses must be conditionally independent. Conditional independence (also called *local independence*) is a basic assumption of most IRT models – including those introduced in this chapter. It means that the probability of responding positively to a particular item is independent of a given respondent's performance on any other item once we have conditioned on (i.e., controlled for) proficiency (θ). That is, given θ , our prediction of how a person will respond to a given item is not aided by knowledge of their performance on any other items; θ encapsulates all we need to know about their proficiency. Conditional independence is testable (e.g., Chen & Thissen, 1997; Rosenbaum, 1988; Yen, 1984, 1993), and when it is violated tends to yield overestimates of the accuracy of estimation (i.e., *estimated* standard errors are too small; Thissen, Steinberg, & Mooney, 1989). Such violations lead to overconfidence because erroneously assuming independence is equivalent to assuming that the information provided by each item is unique when in fact at least some of it is redundant and consequently we end up thinking we know more than we do.

When an assessment or measurement instrument is unidimensional, conditional independence and unidimensionality are equivalent concepts: both are assertions that proficiency describes the complete latent space (Lord & Novick, 1968; Lord, 1980); however, conditional independence requires only that the entire latent space be specified and thus can be a property of more general models as well. Some of these more general models are classified as *multidimensional* because they include more than one trait of interest for each respondent; others are (for practical purposes) still treated as unidimensional but include additional parameters for modeling some of the common dependencies that arise when a subset of items share a common stimulus (e.g., reading passage, clinical scenario). In both cases, the complete latent space is specified by extending the unidimensional models we introduced above, which gives rise to conditional independence. Detailed descriptions of these more general models lie outside the scope of this chapter but curious readers are referred to the references given in Note 2 as well as Wainer, Bradlow, and Wang (2007).

Joint estimation of item parameters and proficiencies

It is reasonable to begin our discussion of joint estimation by simply extending the method of maximum likelihood. Thus, we start by rewriting the likelihood equation shown in Equation 5.7 to include the full matrix of respondents: the likelihood of observing the response matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ from a sample of N independent respondents can be represented as

$$L(\mathbf{X} | \boldsymbol{\theta}, \mathbf{B}) = \prod_n P(\mathbf{x}_n | \theta_n, \mathbf{B}) = \prod_n \prod_i P(x_{ni} | \theta_n \mathbf{B}_i), \quad (5.8)$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)$ and $\mathbf{B} = (a_1, b_1, c_1, \dots, a_I, b_I, c_I)$ are all considered unknown, fixed parameters.¹² Note that Equation 5.8 is simply the product of each respondent's likelihood (from Equation 5.7) – i.e., Equation 5.8 gives the joint likelihood of the full response matrix.

Analogous to the method of maximum likelihood described above, *Joint Maximum Likelihood* (JML) estimates are obtained by finding the values of each θ_n and each β_i that together maximize the likelihood depicted in Equation 5.8. This can be accomplished with respect to $\boldsymbol{\theta}$ and $\mathbf{B}$ jointly

by applying exactly the same ideas discussed earlier in the context of estimating θ ; however, with both Equation 5.7 and Equation 5.8, direct maximization often proves unsatisfactory. Two reasons for this concern us here: lack of consistent estimators and numerical instability.

On consistency

A *consistent* (or asymptotically consistent) estimate is one that converges to its true value as the number of observations increases. When item parameters are known, maximum likelihood estimates of proficiency have the property of consistency; and likewise, so do maximum likelihood estimates of item parameters when proficiencies are known. The problem of consistency arises in the simultaneous estimation of item and respondent parameters, which was first described formally by Neyman and Scott (1948). These authors used the terms *structural* and *incidental* to distinguish between those parameters whose number are unaffected by the number of observations from those that increase with the number of observations, respectively, and they showed that estimates of structural parameters may not be consistent when estimated simultaneously with incidental parameters. In the context of IRT, both item parameters and proficiencies *can* be viewed as incidental parameters (as the number of items or respondents increase, so too does the number of model parameters) and it has been shown by Haberman (1977) for the 2-PL and conjectured by Swaminathan and Gifford (1983) (among others) for the 3-PL that JML estimates of θ and $\mathbf{B}$ are consistent as *both* the number of items and the number of respondents increase without bound. In practice, item parameters for a given test are typically taken to be structural (the number of items is fixed regardless of the number of respondents) whereas proficiencies are taken to be incidental (their number increases with the number of respondents). Under these conditions, JML cannot be assumed to yield consistent estimates.

Conditional Maximum Likelihood (CML) and *Maximum Marginal Likelihood* (MML) estimation may be seen as two strategies for addressing the problem consistency and both succeed by eliminating the incidental parameters from the likelihood.¹³ CML accomplishes this in the context of the 1-PL by replacing θ with its sufficient statistic – its raw score; however, this approach does not extend to more general IRT models wherein sufficient statistics either cannot be used or, in the case of the 3-PL, do not exist. Given these limitations and others, CML is not widely used even for the 1-PL (e.g., WINSTEPS [Linacre, 2011] does *not* use CML) and we do not dwell on it further (the interested reader is referred to Wainer, Morgan, and Gustafsson [1980] for details on this procedure). A more general solution involves integrating the incidental parameters, θ , out of the likelihood, to yield consistent maximum likelihood estimates of the structural parameters, $\mathbf{B}$. This approach was first suggested by Bock and Lieberman (1970) and because it involves estimation of $\mathbf{B}$ in the marginal distribution of θ , it was (eventually) denoted *marginal maximum likelihood estimation* or, more recently (and accurately), the *maximum marginal likelihood estimation*, both of which are referred to as *MMLE* for short.

Maximum marginal likelihood estimation

To integrate θ out of the likelihood, MMLE requires information about the population proficiency distribution that can be incorporated into the IRT model during the marginalization process. This is accomplished using Bayes theorem – although historically MMLE is not considered a Bayesian estimation technique per se (Baker & Kim, 2004). Suppose we have no information that would lead us to have different prior beliefs about respondent traits for different respondents from a given population and further that the density function for this population's proficiency distribution is specified by $g(\theta | \tau)$. For example, the proficiency distribution could be normal

with a mean and variance given by $\boldsymbol{\tau}$. Whatever the specific form of $g(\theta | \boldsymbol{\tau})$ may be, it must be specified *a priori* – although its parameters, $\boldsymbol{\tau}$, can be estimated from previous data, or even from the same data as those from which the item parameters are to be obtained. Consistency follows if maximizing values for $\mathbf{B}$ are obtained after marginalization with respect to $g(\theta | \boldsymbol{\tau})$. That is, maximum marginal likelihood (MML) estimates of $\mathbf{B}$ maximize:

$$L(\mathbf{B} | \mathbf{X}) = \prod_n \int P(\mathbf{x}_n | \theta, \mathbf{B}) g(\theta | \boldsymbol{\tau}) d\theta. \quad (5.9)$$

Numerical procedures for accomplishing MML estimation are described by Bock and Aitkin (1981), Levine (1985), Samejima (1983), and Johnson (2007) among others. Once consistent estimates of the structural parameters, $\mathbf{B}$, are obtained, they can be treated as known for the purpose of estimating the incidental parameters, θ , using the method of maximum likelihood described above (or one of its refinements).

On numerical instability

With the problem of consistency adequately addressed, we are left only with the problem of numerical instability. Numerical instability refers to the failure of a numerical procedure to yield reasonable and finite parameter estimates. Such failures can arise when estimating the item parameters, $\mathbf{B}$, using MMLE and, once these structural parameters are estimated, when estimating proficiency, θ ; however, in both cases, these problems can be avoided by taking further precautions during estimation.

In the case of proficiencies, numerical instability can result from a number of causes. For example, maximum likelihood estimates will not be finite when a respondent answers all items positively (or all negatively). Further, in the 3-PL model some response patterns (e.g., many patterns at below chance levels) also yield infinite proficiency estimates. Also, the likelihood surface does not always yield a single mode; sometimes it can have a number of local extrema. In these cases, solutions to the derivative equations may correspond to a local, but not the global, maximum of L , or even to a local minimum. For these reasons, refinements of the method of maximum likelihood are typically employed when estimating proficiency.¹⁴ The problems of infinite estimates are usually solved by utilizing a prior proficiency distribution; those of local extrema are often resolved through the use of a “good” (i.e., close to the global maximum) starting value for the Newton–Raphson iterations (e.g., one based on a rescaled logit of percent of items scored positive), or running mode-finding algorithms from multiple starting values.

In the case of item parameters, $\mathbf{B}$, additional structure beyond that specified by the IRT model is needed to prevent estimates from drifting out of bounds. This is especially important when the number of respondents or items is few, the response matrix is sparse, respondents are homogeneous with respect to the latent trait of interest, items are not well-targeted to respondents (e.g., items are too easy), or in general whenever the 3-PL is used. Incorporating prior distributions for item parameters into the model using Bayes’ theorem leads to an extension of MMLE called *marginalized Bayesian estimation*,¹⁵ which effectively prevents item parameter estimates from taking on unreasonable values (Mislevy, 1986).

On the accuracy of proficiency estimates

So far we have been concerned with obtaining a point estimate of a respondent’s proficiency. The point estimate we have adopted is the most likely value: the mode (maximum) of the likelihood function. But even a quick glance at the bottom graph in Figure 5.6 tells us that there is

a substantial likelihood of other values, at least with a two-item test. The standard deviation of the (asymptotically normal) likelihood distribution is commonly used to characterize the precision of the proficiency estimate. (Typically, this distribution is sufficiently normal for most purposes when the assessment is longer than 20 items (Samejima, 1977). If this distribution is narrow, then we are assured that the proficiency estimate we provide is a good one. If the likelihood is broad, we are less sure. In practice we can increase the accuracy of an IRT estimate of proficiency by increasing the length of the test; however, if we add an item that is much too easy, much too hard, or not very discriminating for a given respondent, such an item's $P(\theta)$ and $Q(\theta)$ will be essentially constant in the neighborhood of the likelihood mode – and multiplying the likelihood by a constant will not shrink its standard deviation.

When the number of items administered to a respondent is large, the variance of the respondent's likelihood can be approximated as the reciprocal of the Fisher *information function*,

$$I(\theta) = \sum_i \frac{P'_i(\theta)^2}{P_i(\theta)Q_i(\theta)}, \quad (5.10)$$

where $P'_i(\theta)$ is the first derivative of $P_i(\theta)$ with respect to θ . Thus, the standard deviation of the likelihood, called the *standard error of estimation*, is given by:

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}. \quad (5.11)$$

In practice θ is unknown and thus $I(\hat{\theta})$ is typically substituted for $I(\theta)$ in Equation 5.11. An important feature of Equation 5.11 is that the standard error is conditioned on proficiency, which we observed above is typically not the case in the true score theory paradigm. This is an appealing property because it is reasonable to expect our confidence in respondents' proficiency estimates to vary depending on the quality (e.g., discriminating power) of the items we present them and how well-matched items are to respondents' proficiency levels.

Item and test information

The information function, Equation 5.10, has the attractive features of being additive over items and being independent of the item responses. Thus, while the information function given in Equation 5.10 led to the important result shown in Equation 5.11, we see that it also provides a way to quantify the contribution of information of each item for any given θ as well as to summarize the information for a set of items (such as a test). Figure 5.7 shows a typical item information function (Figure 5.7a); the item information functions for a 15-item test (Figure 5.7b); the sum of these 15 item information functions, which yields the associated test information function (Figure 5.7c); and the reciprocal of the square root of the test information function, which yields the associated standard error of estimation function (Figure 5.7d).

Item information functions, test information functions, and standard error of estimation functions can be profitably used to optimize certain measurement objectives during test assembly (e.g., Lai et al., 2011). For example, depending on the inferences being drawn from test scores, test designers may decide upon a maximum standard error that can be tolerated. A target test information function can then be specified that corresponds to standard errors equal or less than the given maximum within the proficiency range of interest to the examiners. Given a sufficient pool of calibrated items, items can be selected during test assembly such that the target is met. In other cases, examiners may wish to classify respondents based

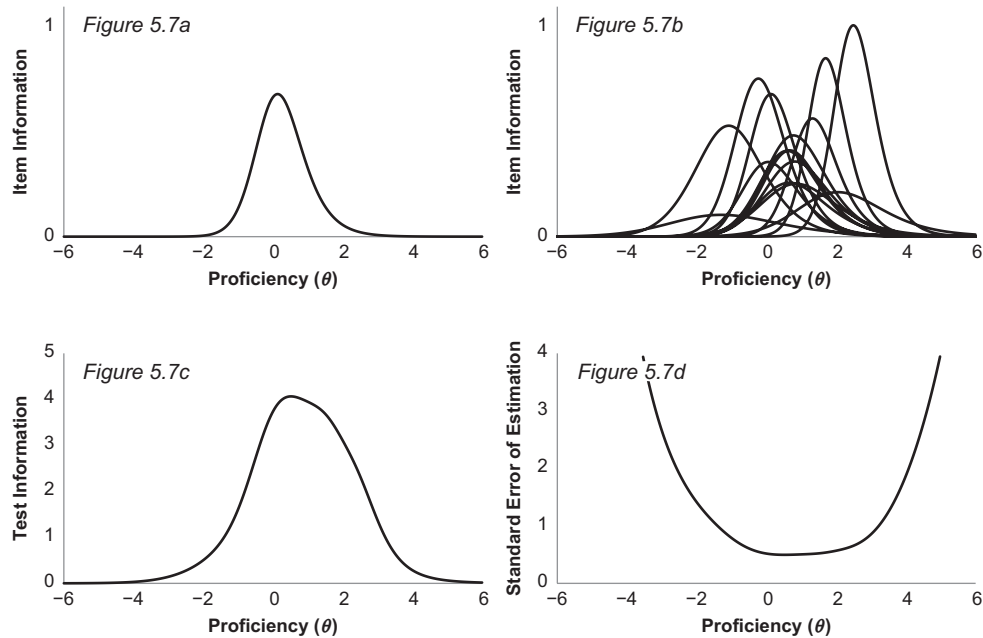


Figure 5.7 An item information function (Figure 5.7a); item information functions for a 15-item test (Figure 5.7b); the associated test information function (the sum of item information functions from Figure 5.7b; Figure 5.7c); the associated standard error of estimation function (the reciprocal of the square root of the test information function; Figure 5.7d).

on their trait level relative to a given cut-score. When this is the goal, classification errors can be reduced by selecting items that maximize test information around the cut-score. Note, however, that whenever items are selected on the basis of their item parameter *estimates*, the problem of selection bias arises. Practitioners should take steps to control for errors of this kind (Hambleton & Jones, 1994).

Evaluating model-data fit

IRT models offer many attractive features; however, IRT is based on strong assumptions and when the data do not fit the model, the advantages of these models may not be realized. Most commercial IRT estimation software packages provide some measure of model-data fit; although, thoughtful consideration of their appropriateness and utility for each particular measurement problem is recommended because no single statistic works well under all conditions and none are universally accepted. Most fit indices are χ^2 statistics of one kind or another (e.g., Yen's, 1981, Q1), but alternatives have also been proposed (e.g., Orlando & Thissen's, 2000, 2003, $S - \chi^2$ and Glas's, 1998, 1999, Lagrange Multiplier procedure). Another perhaps more intuitive approach is to compare the observed response data to the estimated model by examining residuals and standardized residuals (see Hambleton, Swaminathan, & Rogers, 1991, for an introduction to this type of analysis). An excellent review of these and other strategies for assessing model-data fit can be found in Swaminathan, Hambleton, and Rogers (2007).

A particular failure of model-data fit that deserves comment is when item parameters vary across respondent subgroups such as gender, race/ethnicity, or diseases (e.g., Teresi et al., 2009).

It follows from the model property of parameter invariance that group membership should not affect the probability of a given response once we control for proficiency. When group membership *does* influence the probability of a given response for respondents of equal proficiency, it is given a special name: *differential item functioning* or simply *DIF*. Because it is generally undesirable and often unacceptable for group membership to affect respondent performance, DIF detection is an important subarea in assessment.

With moderate to large samples, DIF analyses may be conducted using observed score methods (see Holland & Wainer, 1993, for a review of these methods); however, when sample sizes are small, a strong model is required and in situations like this IRT-based methods can be useful. Perhaps the three best-known IRT-based strategies for identifying DIF involve (i) comparing model-data fit using a likelihood ratio test between two nested IRT models – one that estimates group-specific common item parameters and one that estimates combined-group common item parameters (Thissen, Steinberg, & Wainer, 1988), (ii) comparing the common item parameters directly using a chi-squared test (Lord, 1980), or (iii) comparing the signed (or unsigned) area between each group's ICC for each suspected item with the expected sampling distributions of areas assuming invariance (Raju, 1990). In addition to these primary sources, readers are again referred to Holland and Wainer's (1993) book for further discussion of these approaches.

Estimation software

Once response data are collected, the IRT model must be estimated using special-purpose software. Readers comfortable with R (R Development Core Team, 2008) can find a number of open-source libraries that perform IRT analyses. Other software has been made available free-of-charge by the authors – e.g., SCORIGHT (Wang, Bradlow, & Wainer, 2004), ICL (Hanson, 2002), and BIGSTEPS (Wright & Linacre, 1997). Additionally, many commercial software packages exist that estimate IRT models; Table 5.2 lists five that are well known.

Table 5.2 Some well-known commercial software packages for estimating IRT models

<i>Software Package</i>	<i>Type of Response Data</i>		<i>Comments</i>
	<i>dichotomous</i>	<i>polytomous</i>	
BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003)	X		1-PL, 2-PL, 3-PL
PARSCALE (Muraki, & Bock, 2003)	X	X	1-PL, 2-PL, 3-PL, GRM, and others*
IRTPRO (Cai, Thissen, & du Toit, 2011)	X	X	1-PL, 2-PL, 3-PL, GRM, and others*
Xcalibre (Guyer & Thompson, 2011)	X	X	1-PL, 2-PL, 3-PL, GRM, and others*
WINSTEPS (Linacre, 2011)	X	X	1-PL and others*

Note: Models introduced in this chapter can be estimated using these packages.

* As noted above, although outside the scope of this chapter, a variety of alternative polytomous IRT models exist. Several of these can be estimated using these and other software packages.

Conclusion

IRT developed out of the need to provide a more coherent connection among the disparate pieces that make up a typical test or survey instrument. In this sense, IRT can be seen as a connective theory – in which the test item is the fungible unit – that follows from Wundt’s sensitivity curves in psychophysics and dosage response curves in epidemiology. Nevertheless, a crucial difference makes IRT’s intellectual and technical contribution special. As with IRT, both sensitivity curves and dosage response curves report the probability of response on the vertical axis; however, while in psychophysics the horizontal axis refers to the *observed* physical magnitude and in dosage response curves, the horizontal axis refers to the *observed* dosage, in IRT the variable on the horizontal axis, proficiency, is *not* observed, it is latent and it must be estimated.

IRT made its first major contribution into healthcare when a very general IRT model was used to calculate skeletal age (Roche, Wainer & Thissen, 1975).¹⁶ The so-called RWT method used a mixture of indicators of maturity that were binary, categorical, ordered, or continuous and put them together with maximal efficiency within an IRT model. This approach was an attractive alternative to the atlas-based methods popular at the time, for in addition to providing an estimate of the child’s skeletal age, it also yielded its standard error. In this way, the estimate of skeletal age was put on a firm statistical footing for the first time.

In the early 1990s, IRT was taken up enthusiastically by researchers interested in developing indices of health outcomes and quality of life. Early pioneers taking this approach were John Ware and his associates (Wu et al., 1991) but researchers continue to make advances in this area (e.g., Anatchkova et al., 2012). With development and distribution of the Patient Reported Outcome Measurement Information System (www.PROMIS.gov), IRT-based measurement will be used increasingly in clinical practice and research in nursing and the health sciences (Aletaha, 2010; Henly, 2010). Thus, while the potential for further applications of IRT in these fields remain considerable, to realize this potential users must have some understanding of the strengths and weaknesses of IRT and IRT methodology. This chapter was meant to provide readers with a starting point for developing such an understanding.

Notes

- 1 While IRT describes respondent behavior probabilistically, some debate exists regarding the precise interpretation of probability in this context. These interpretations come in two flavors: assertions that probabilities arise from (i) sampling a group of respondents, or (ii) variability within an individual respondent (what Holland (1990) calls the ‘stochastic subject’ rationale). The group interpretation assumes that when confronted with an item an individual will respond positively or negatively (in the context of testing, this generally means *correctly* or *incorrectly*). The probability arises from the idea that if a *group* of respondents with a given proficiency responded to an item, some of them will respond positively and the balance will not.

The latter notion – the ‘probability-in-the-soul’ idea – stems from quite a different conception. It asserts that a given person *alone* provides a positive response to a given item with a given probability. This point of view is left over from the psychophysical origins of test theory, in which an experimental subject could be presented with a very weak stimulus (e.g., a dim light, a barely audible sound, a very light touch) and asked if they perceived (e.g., heard, saw, felt) it. Sometimes they would and sometimes they would not. A given stimulus strength that was perceived by the subject 50% of the time was called the “threshold strength” of the stimulus. This is directly analogous to item difficulty, and so it is easy to see how the conception of building the stochastic variability into the person within test theory was a natural outgrowth of the past; however, there are many assessment contexts where it is difficult to reconcile this interpretation with our beliefs about individual respondent behavior.

Historically, these competing interpretations of probability have led to different approaches for estimating model parameters. Readers interested in further details on how various estimation methods rely

on different interpretations of probability are referred to Paul Holland's (1990) presidential address to the Psychometric Society. Here, let us simply note that while probability-in-the-soul may sometimes be a useful way to think about respondent behavior (for example, it is hard to imagine how models that include guessing would have evolved without such a conception), most psychometricians believe the sampling interpretation is a better description of the phenomena encountered in testing and that claims of sample-free item analysis should not be taken literally.

- 2 Although they are much less widely used, multidimensional IRT models do exist (e.g., Reckase, 1985, 1997; McDonald, 2000; Segall, 2001).
- 3 Although it is by far the most widely used, the logistic is not the only kind of response function employed for joining ability and difficulty; several different kinds of ogive functions have been used. The cumulative normal is a popular alternative, but even monotone splines have proved to be sometimes useful (Ramsay, 1991, 1992).
- 4 Although the Rasch Model and the 1-PL are in some formal sense identical, certain conventions characterize the Rasch Model's use. These are discussed later in the section on indeterminacy.
- 5 While there is widespread agreement that the 1-PL has a number of unique properties that deserve thoughtful consideration, the benefits of these properties is not agreed upon. These disagreements follow from more fundamental differences in theoretical perspectives between self-described followers of Rasch and users of more general models. Moreover, the *practical* importance of these differences in perspectives is itself another source of disagreement – although some, perhaps especially proponents of the Rasch Model, seem to view it as very important indeed (e.g., Andrich, 2004). In any case, while outside the scope of this chapter, we take the view that understanding these perspectives is a worthwhile and enriching endeavor; however, curious readers are cautioned that publications in this area are sometimes polemics rather than objective comparisons.
- 6 It is not uncommon to see D substituted for a in Equations 5.2 and 5.3, where D is a scaling factor equal to $D = 1.7$. In so doing, the more mathematically tractable logistic formulation closely resembles the normal ogive function, which preceded the use of the logistic function in the historical development of IRT. If used, D should be viewed merely as a convenience, which like any scaling convention can simplify interpretation for users accustomed to parameters being scaled in a particular manner. This convention should *not* be taken to mean that the normal form has greater intrinsic merit (Birnbbaum, 1968).
- 7 The slope of Equation 5.2 is given by its first derivative, $aP(\theta)(1 - P(\theta))$.
- 8 In the case of the GRM, we deviate from conventional notation somewhat in the hopes of making the presentation clearer to novice users of IRT. For example, it is more common to see $P_k^*(\theta)$ rather than $P_{\geq k}(\theta)$, but it is our view that the latter formulation is more descriptive. There are no occasions where these differences are substantive. For a discussion of the various notation conventions for polytomous IRT models, see Nering and Ostini (2010).
- 9 Recall the absence of a discrimination parameter in Equation 5.1. This absence is problematic if we wish to follow the common (but arbitrary) convention of scaling proficiency such that it has a standard deviation of 1 because for Equation 5.1: $(\theta^* - b^*) \neq (\theta - b)$ except in the case of $\gamma = 1$. For this reason, it is preferable to define the 1-PL in the same manner as the 2-PL – i.e., as having a discrimination parameter, a , in the logit, $a(\theta - b)$ – but with the added constraint that a common value for a be estimated for (or assigned to) all items. In so doing, we can scale proficiency, difficulty, or discrimination in any way we deem convenient. Moreover, this approach highlights the distinction between the 1-PL and the Rasch Model: the latter denotes the *convention* of scaling $a = 1$ – not to any formal difference between the Rasch Model and the 1-PL.
- 10 Equation 5.7 yields the conditional probability of a specified response pattern; however, when $\mathbf{x}_n$ is an actual observed response vector, the term *likelihood* is preferred over *probability*.
- 11 Unlike the example shown in Figure 5.6, in practice respondents typically respond to more than 2 items. This leads to practical problems when computing likelihoods – e.g., when a series of numbers that are bound by 0 and 1 (as $P(\theta)$ and $Q(\theta)$ are) are multiplied, the resultant product rapidly become impractically small. Problems like this can be avoided by transforming the likelihood using logarithms because log-likelihoods and likelihoods are monotonically related and thus share the same maximum. Unlike likelihoods, log-likelihoods do not get impractically small because the logarithm of a product is equal to the sum of the logarithms of its multiplicands ($\ln xy = \ln x + \ln y$). Logarithms have other convenient properties too (e.g., $\ln x^\gamma = \gamma \ln x$) that simplify finding the maximum log-likelihood using numerical methods. Because of these conveniences, the following equation is almost always substituted for Equation 5.7 in practice:

$$\ln L(\mathbf{x}_n | \theta_n, \mathbf{B}) = \sum_i [x_{ni} \ln P_i(\theta_n) + (1 - x_{ni}) \ln Q_i(\theta_n)]$$

- 12 The continued product over items for each respondent is understood to run over only those items administered to that respondent, assuming that the responses to the items not administered can be considered ignorable in the sense of Little and Rubin (2002).
- 13 Paul Holland once remarked that CML should be renamed NML for Not Maximum Likelihood and that MML should be renamed ML for Maximum Likelihood. (Holland's quip was reported by David Thissen in his invited address accompanying receipt of the NCME Career Achievement Award in Chicago, April 18, 2015.)
- 14 While we do not discuss them here, in some cases, researchers in the past have advocated using various ad hoc rules to overcome estimation problems. For example, in the face of zero or perfect scores, Wright and Stone (1979, p. 61) instruct practitioners to "edit the binary data matrix of person-by-item responses such that no person has a zero or a perfect score and no item has a zero or a perfect score."
- 15 Bayesian estimators also exist that do not rely on marginalization of any kind (e.g., Swaminathan & Gifford, 1982, 1985, 1986).
- 16 This research was later extended and improved (Roche, Chumlea, & Thissen, 1988).

References

- Aletaha, D. (2010). From the item to the outcome: The promising prospects of PROMIS. *Arthritis Research and Therapy*, 12, 104.
- Anatchkova, M. D., Rose, M.S.F., Ware, J. E., Jr., & Bjorner, J. B. (2012). Development of an item bank and computer adaptive test for role functioning. *Quality of Life Research*, 21, 1625–1637.
- Andrich, D. (2004). Controversy and the Rasch model. A characteristic of incompatible paradigms? *Medical Care*, 42 (1 Suppl.), I7–I16.
- Baker, F.B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring a respondent's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (chaps. 17–20). Reading, MA: Addison-Wesley.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 46, 443–459.
- Bock, R. D., Lieberman, M. (1970). Fitting a response curve model for dichotomously scored items. *Psychometrika*, 35, 179–198.
- Cai, L., Thissen, D., & du Toit, S.H.C. (2011). *IRTpro for Windows* [Computer software]. Lincolnwood, IL: Scientific Software.
- Chen W.-H., & Thissen D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265–289.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.
- Fries, J. F., Cella, D., Rose, M., Krishnan, E., & Bruce, B. (2009). Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. *Journal of Rheumatology*, 36, 2061–2066.
- Glas, C.A.W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, 8, 647–667.
- Glas, C.A.W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika*, 64, 273–294.
- Gómez, J., Hidalgo, M. D., & Tomás-Sábado, J. (2007). Using polytomous item response models to assess death anxiety. *Nursing Research*, 56, 89–96.
- Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley. (Reprinted in 1987 by Lawrence Erlbaum Associates: Hillsdale, NJ.)
- Guyer, R., & Thompson, N. A. (2011). *User's manual for Xcalibre item response theory calibration software, version 4.1.3*. St. Paul, MN: Assessment Systems.
- Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics*, 5, 815–841.
- Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education*, 7, 171–186.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hanson, B. A. (2002). *IRT Command Language (ICL)* [Computer software]. Available at www.b-a-h.com/software/irt/icl/
- Hays, R., Liu, H., Spritzer, K., & Cella, D. (2007). Item response theory analyses of physical functioning items in the medical outcomes study. *Medical Care*, 45(5 Suppl.), S32–S38.
- Henly, S. J. (2010). The promise of PROMIS [Editorial]. *Nursing Research*, 59, 77.
- Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, 55, 577–601.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software*, 20. Retrieved from www.jstatsoft.org/v20/i10/paper
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.
- Lai, J. S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., & Stone, A. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation*, 92(10 Suppl.), S20–S27.
- Levine, M. (1985). The trait in latent trait theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 41–65). Minneapolis: Computerized Adaptive Testing Laboratory, Department of Psychology, University of Minnesota.
- Linacre, J. M. (2011). *Winsteps Rasch Measurement* [Computer software]. Available from www.winsteps.com
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.
- Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 61, No. 4.
- Lord, F. M. (1952). *A theory of test scores*. (Psychometric Monograph, No. 7). Richmond, VA: Psychometric. Retrieved from www.psychometrika.org/journal/online/MN07.pdf
- Lord, F. M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 18, 57–75.
- Lord, F. M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13, 517–548.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.
- McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, 24, 99–114.
- Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, 51, 177–195.
- Muraki, E., & Bock, R. D. (2003). *Parscale (Version 4)* [Computer software and manual]. Lincolnwood, IL: Scientific Software.
- Nering, M., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge.
- Neyman, J., & Scott, E. L. (1948) Consistent estimates based on partially observed observations. *Econometrica*, 16, 1–32.
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50–64.
- Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27, 289–298.
- R Development Core Team (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org
- Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14, 197–207.
- Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56, 611–630.
- Ramsay, J. O. (1992). *TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data*. (Technical Report). Montreal, Quebec: McGill University.

- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmark's Paedagogiske Institut.
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401–412.
- Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25–36.
- Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Cella, D. (2007). PROMIS cooperative group: Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45, S22–S31.
- Roche, A. F., Wainer, H., & Thissen, D. (1975). *Skeletal maturity: The knee joint as a biological indicator*. New York, NY: Plenum.
- Roche, A. F., Chumlea, W. C., & Thissen, D. (1988). *Assessing the skeletal maturity of the hand-wrist: Fels method*. Springfield, IL: Charles C. Thomas.
- Rosenbaum, P. R. (1988). A note on item bundles. *Psychometrika*, 53, 349–360.
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from www.psychometrika.org/journal/online/MN17.pdf
- Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement*, 1, 233–247.
- Samejima, F. (1983). Some methods and approaches of estimating the operating characteristics of discrete item responses. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 159–182). Hillsdale, NJ: Lawrence Erlbaum.
- Segal, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, 66, 79–97.
- Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, 7, 175–192.
- Swaminathan, H., & Gifford, J. A. (1983). Estimation in the three-parameter latent trait model. In D. Weiss (Ed.), *New horizons in testing* (pp. 13–29). New York, NY: Academic Press.
- Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, 50, 349–364.
- Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, 51, 589–601.
- Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 683–718). Amsterdam: North Holland.
- Teresi, J. A., O'cepek-Welickson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Jones, R. N., . . . Cella, D. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychological Science Quarterly*, 51, 148–180.
- Thissen, D. (April 18, 2015). Item Response Theory, Serendipity, and Bad Questions. Invited Address at the annual meeting of NCME Chicago, Ill.
- Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, 26, 247–260.
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–170). Hillsdale, NJ: Lawrence Erlbaum.
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.
- Wainer, H. (2011). The first step toward wisdom. *CHANCE*, 24(2), 60–61.
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.
- Wainer, H., Morgan, A., & Gustafsson, J.-E. (1980). A review of estimation procedures for the Rasch model with an eye toward longish tests. *Journal of Educational Statistics*, 5, 35–64.
- Wang, X., Bradlow, E. T., & Wainer, H. (2004). *SCORIGHT: A computer program for scoring tests built of testlets including a model for covariate analysis (Version 3.0)* [Computer program]. Princeton, NJ: Educational Testing Service.

- Wright, B. D., & Linacre, J. M. (1997). *BIGSTEPS Rasch analysis* [Computer software]. Chicago, IL: MESA Press.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.
- Wu, A. W., Rubin, H. R., Mathews, W. C., Ware Jr., J. E., Brysk, L. T., Hardy, . . . Richman, D. D. (1991). A health status questionnaire using 30 items from the Medical Outcomes Study. Preliminary validation in persons with early HIV infection. *Medical Care*, 29, 786–798.
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245–262.
- Yen, W. M. (1984). Effect of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125–145.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187–213.
- Yount, S. E., Choi, S. W., Victorson, D., Ruo, B., Cella, D., Anton, S., & Hamilton, A. (2011). Brief, valid measures of dyspnea and related functional limitations in chronic obstructive pulmonary disease (COPD). *Value Health*, 14, 307–315.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG* [Computer software]. Lincolnwood, IL: Scientific Software.

6

BEHAVIORAL OBSERVATION

Tondi M. Harrison

Direct behavioral observation is used to collect data describing behaviors that unfold over time. This powerful research approach is used to study health experiences across the lifespan, especially among individuals who cannot speak for themselves, including infants and young children, persons with mental impairments and disabilities, elderly with dementia, and critically ill patients. Basic principles and strategies can be extended to observation of social interactions among dyads, such as patients and caregivers, or among members of larger groups, such as families. Direct observation is used in laboratory and naturalistic settings such as homes, hospital units, and schools. Direct observation in finely grained time provides detailed information about change in a behavior and behavioral sequences over time. In intervention studies, serial measurements obtained via direct observation provides information about mechanisms effecting change.

The purpose of this chapter is to provide an overview of direct behavioral observation methods. Issues addressed are (a) definition and comparison with other methods of assessing behavior; (b) indications for use of behavioral observation; (c) principles for direct behavioral observation, including methods and materials for observing and recording behavior; and (d) assessment of data quality.

Definition

Direct behavioral observation is the systematic observation of behavior as it occurs in real-time. Observations may be made by those enacting the behaviors (self-observation) or by trained observers present in the behavioral setting or reviewing electronic data recordings at a later time.

Direct behavioral observation is distinct from self-report or informant approaches used for indirect assessment of behavior (Shapiro & Kratochwill, 2000). Self-report methods involve participants supplying their own responses to questions or statements (e.g. interviews or questionnaires) pertaining to behaviors, emotions, or cognitions that occurred in other places and at other times (Cone, 1978). Informant reports are gathered from individuals with personal experience with the study participant such as teachers, parents, spouses, or health care providers (Merrell, 2000). Informant ratings provide perceptions of an individual's behavior based on the rater's experience rather than reports of actual behavior as it is occurring (Merrell, 2000).

Both indirect and direct behavioral assessment offer unique perspectives on the individual, dyad, or group from the identified population under study. Indirect observation using rating

systems may better capture global behavioral characteristics or tendencies, whereas direct behavioral observations are more precise for examining specific behaviors and small behavioral changes, as well as antecedent and consequent behaviors (Snyder et al., 2006).

Indications

Direct behavioral observation is the method of choice for research examining behavior or behaviors with overt manifestations that are expected to occur or to change over small units of time or within short time frames (Snyder et al., 2006). Interpersonal interactions (dyadic, triadic, or group), behavioral change or stability before and after a discrete nursing or medical intervention or event, or developmental changes in specific health-related behaviors over time are all indications for use of direct observational methods. Table 6.1 provides a list of research studies utilizing a variety of direct behavioral observation methods in a number of different settings and populations.

Table 6.1 Studies using behavioral observation

<i>Study^a</i>	<i>Who (Population)</i>	<i>What (Behavior Observed)</i>	<i>Where (Setting)</i>	<i>When (Time Sampling)</i>	<i>How (Materials)</i>
Campbell, 2010	Adults with respiratory distress	Restlessness, accessory muscle use, grunting, nasal flaring, facial expression	Hospital	Momentary time sampling	Pencil and paper
Clayman, 2012	Women with breast cancer	Communication: shared decision making	Oncology clinic	Continuous	Video recording
de Lemus, 2012	Adult men and women	Posture and facial expressions	Laboratory	Momentary time sampling	Video recording
Dykstra, 2013	Children with autism	Child and adult vocalizations and conversational turns	Preschool classroom	Continuous	Audio recordings with LENA system
Harrison, 2009	Infants after open heart surgery	Maternal infant interaction during feeding	Hospital and home	Continuous	Video recording
Hayes, 2008	Community dwelling adults	Movement, activity, and walking speed in the home	Home	Continuous	Motion sensors
Heathcock, 2004	Preterm and full-term infants	Kicking	Home	Continuous	Video recording
Held, 2005	Anesthesiologists	Anesthesia work processes	Operating room suite	Continuous	Flexible Interface Technique for complex behavioral processes
Jovanov, 2005	Adults	Physical activity and physiologic measures	Home	Continuous	Wireless Body Area Network

Table continues

<i>Study^a</i>	<i>Who (Population)</i>	<i>What (Behavior Observed)</i>	<i>Where (Setting)</i>	<i>When (Time Sampling)</i>	<i>How (Materials)</i>
June, 2012	Adult smokers	Smoking topography	Laboratory and home	Continuous	Portable device to measure smoking topography (CReSSmicro)
Kendler, 2008	Reared-apart twins	Facial expressions (Facial Action Coding system)	Laboratory	Continuous	Video recording
Lavelle, 2013	Schizophrenic adults	Gestures and head nodding during interaction	Laboratory	Continuous	Video recording and Vicon motion-capture system
Loman, 2013	Post-institutionalized 10- and 11-year-old children	Performance on computer stimulus-response test	Laboratory	Continuous	Computer task
Meeks, 2011	Depressed elderly nursing home residents	Activity, mood, and engagement	Nursing home	Continuous	Video recording
Mesman, 2008	Children 1–4 years of age	Physically aggressive behaviors	Laboratory	Continuous	Video recording
Mitchell, 2009	Families of infants and toddlers with cystic fibrosis	Mealtime interactions	Home	Continuous	Video recording
Nettle, 2011	Two socioeconomically disparate neighborhoods	Smoking, drinking alcohol, exercising (running)	Community	Continuous	Observations recorded in real-time on digital voice recorder
Reed, 2011	Adults using recreational trail	Physical activity behaviors	Community	Momentary time sampling	Pencil & paper
Robbins, 2011	Female adults with rheumatoid arthritis	Sighing and depression	Home	Momentary time sampling	Electronically activated audio recorder
Roberts, 1991	Adults in self-help group	Helping behaviors	Meeting room	Continuous	Pencil and paper
Silk, 2011	7–17-year-olds with depression	Emotion, activity, location, and social companion	Community	Momentary time sampling	Cell phone prompt
Thomas, 2011	Adults with alcoholism	Alcohol consumption	Laboratory	Continuous	Video recording
Thoyre, 2012	Premature infants	Behavior during feeding	Neonatal Intensive Care Unit	Continuous	Audio and video recordings
van Oorsouw, 2011	Staff members caring for clients with intellectual disabilities	Non-verbal and verbal expressions of emotion	Inpatient mental health unit	Partial interval recording	Video recording
Vázquez, 2011	Ventilated adults in ICU	Behavioral indicators of pain	Hospital ICU	Momentary time sampling	Pencil and paper

^aFirst author's last name is listed.

Observing and recording data

Critical initial steps in designing behavioral observation studies include: (a) developing mutually exclusive categories of relevant behaviors, (b) defining explicit criteria for each behavioral category, and (c) defining the observational context, including persons, place, and time (see Figure 6.1; Snyder et al., 2006). A variety of methods and materials are available to support accuracy and consistency of behavioral observation in many research settings.

Mutually exclusive behavioral categories

Selecting the appropriate behavior to observe may be the single most important decision made in designing studies that include direct behavioral observation. Initially, a narrative is used to summarize thoughts and ideas about the phenomenon of interest (Skinner, Rhymer, & McDaniel, 2000). These ideas may arise from clinical practice, be prompted by findings in the research literature, or be based on theory. These exploratory ideas about the target behavior are broad in order to ensure that the universe of relevant content represented as general behavioral descriptions of the concept are developed and specific characteristics are listed. From these narratives, hypotheses suggest and direct development of pertinent categories of behaviors for observation, as well as identify those that are irrelevant (Skinner et al., 2000). These categories are known as molar behaviors (Suen & Ary, 1989) and are described as high level, abstract, latent phenomenon that are not directly observable, but which direct and lead to operational definitions of the observable behaviors of interest (see Table 6.2).

Explicit criteria

While molar behaviors describe the overarching concept, the fundamental units of observation are specifically chosen manifestations of molar behaviors known as molecular behaviors (see Table 6.2; Suen & Ary, 1989). An observational protocol often uses multiple molecular behaviors to measure a molar behavior. Identifying and defining explicit criteria for the molecular behaviors of the phenomenon of interest is critical to assuring content validity of the study and helps determine the optimal observational tool. In existing instrumentation and newly developed

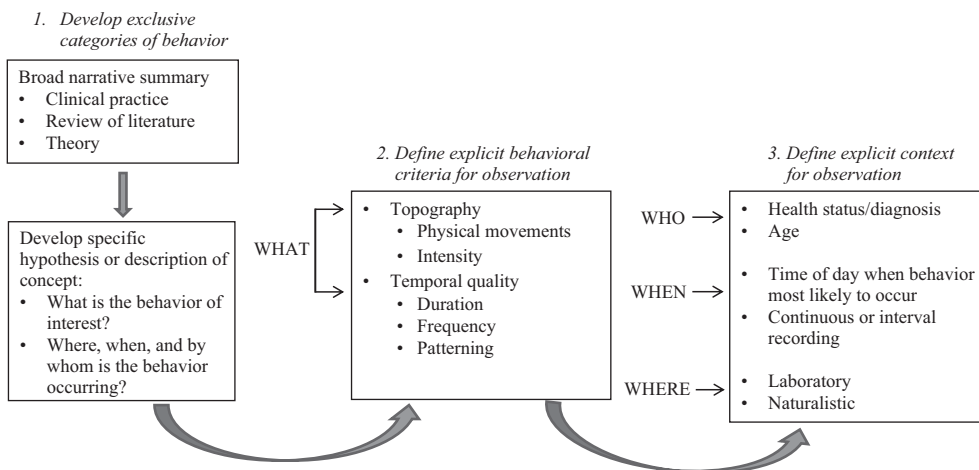


Figure 6.1 Three initial steps in designing behavioral observation studies.

Table 6.2 Differentiating molar concepts and molecular behaviors

<i>Exemplar Molar Concept</i>	<i>Operational Definition of the Molar Concept</i>	<i>Molecular Behaviors to Be Measured</i>
Aggression	Behavior that is aimed at and may cause harm to people, animals, or objects, and is not because of motor limitations, or part of age-appropriate play and exploration.	Kicking, hitting, biting, pinching, scratching, shaking, pushing, stamping, throwing, and physically threatening to perform any of these behaviors (Mesman et al., 2008)
Infant Learning	Increased frequency of kicking with structured task compared to baseline.	A kick is defined as >15 degree extension of the hip and knee with immediate recoil into flexion (Heathcock et al., 2004)
Respiratory Distress	The physical and or emotional suffering that results from dyspnea (the subjective awareness of altered or uncomfortable breathing).	Use of accessory muscles, restlessness, nasal flaring, grunting at end-expiration, fearful facial expression (Campbell, Templin, & Walch, 2010)
Quality of Parental Interaction (expressed in tone of voice)	The extent to which parent's voice lacks inflection, expressiveness, or range of affect.	1 = Very flat or constricted tone of voice is characteristic; no emotion; monotonic 2 = Flat or constricted tone of voice is predominant; brief, fleeting periods of emotion in voice 3 = Some emotion, inflection, or change in pitch is present; tone is flat or constricted much of the time 4 = Brief or fleeting periods of flatness or constriction; affective range in voice is usual 5 = No flatness or constriction in voice; characteristically expressive (Clark, 2010)

protocols, molecular behaviors are based on identifiable topographies and temporal characteristics of the behavior (Skinner et al., 2000).

Topography of behavior

Topography is the shape of a behavior as it is expressed in time (i.e., the physical movements or sounds associated with a behavioral phenomenon; Skinner et al., 2000). For example, the topography of pain behaviors may include physical changes in facial muscles, such as frowning, lowering of eyebrows, or tight closing of eyelids, as well as alterations in vocalizations, such as moaning or crying (Vázquez et al., 2011). The topography of a behavior may also include intensity, i.e., within the same physical movements involved in a behavior, the magnitude or quality with which the behavior is expressed may differ. For example, a topography of smoking includes the mechanics of inhalation, the intensity of which can be quantified by measuring puff volumes (June et al., 2012). The observable behaviors comprising the topography need to be specifically described and may include, as appropriate, numerical values of levels of intensity (see Table 6.2). Language used in the topographical descriptions must clearly describe an observable behavior; language requiring an observer to make a judgment about the participant's emotional state or intentions (e.g., depressed, angry, happy, determined) should be avoided (Suen & Ary, 1989).

Temporal characteristics

Behavior, whether as individual actions or interpersonal processes, occurs over distinct periods of time. The study of behavior requires attention to specific temporal parameters of interest. These parameters may include duration of behavior (how long the behavior lasts), frequency of behavior (how often the behavior occurs), and patterning of behavior (time between behaviors or time between a stimulus and a behavioral response; McGrath & Tschan, 2004; Suen & Ary, 1989).

A behavioral duration is measured as the number of units of time between the beginning and the end of the behavior of interest (McGrath & Tschan, 2004). For example, researchers have defined duration of infant feeding as the time between the first insertion of the nipple into the infant's mouth and the final removal or release of the nipple from the infant's mouth (Thoyre, Holditch-Davis, Schwartz, Roman, & Nix, 2012). Duration does not pertain to targeted behaviors that are discrete and not timed (e.g., a cough or a sigh).

Frequency refers to the number of times a targeted behavior occurs during a predefined period of time. In order to determine a frequency of behavior, definitions include how frequencies will be counted, i.e., what behaviors or sequence of behaviors constitute one episode. For example, in a study of physical aggression in preschoolers, consecutive identical behaviors were coded as discrete occurrences if they were separated by at least two seconds (Mesman et al., 2008).

Patterning of behavior is measured by the spaces between behavioral events, such as the time between antecedent and consequent behaviors or between a stimulus and a response (McGrath & Tschan, 2004). Observations of temporal patterning of behavior may include either precise measurement of units of time or qualitative assessments. For example, behavioral response to a stimulus can be precisely measured in units of time with laboratory-based computer tasks (Loman et al., 2013). Qualitative assessments of temporal behavioral patterns are used in the Parent-Child Early Relational Assessment tool, which includes items such as measuring parental contingent responsivity, i.e., how quickly and frequently a parent responds to the child's actions or cues (Clark, 1985; 2010).

Observational context

Defining the context is a critical component in planning a behavioral observation (see Figure 6.1). The context includes descriptions of who will be observed, where the observations will take place, and when the observations will occur.

Who

Behavioral observations can be utilized for data collection on individuals of any age and any health status. Units of persons can also be observed, such as mother-infant dyads (Harrison, 2009), families (Mitchell et al., 2009), groups (Roberts et al., 1991), and communities (Nettle et al., 2011). At a minimum, the individuals or groups to be observed need to be defined by age and health status or diagnosis. Additional criteria will be based on the research question of interest.

Where

Behaviors can be observed in both laboratory and naturalistic settings. When carefully chosen and constructed, either setting can provide valid data generalizable to the population and environment of interest.

Laboratory settings allow control of potential confounding external environmental stimuli, so that each study participant has a similar experience (Hintze, Stoner, & Bull, 2000). Laboratory

observations and controlled experiments are also used to elicit behavior that would occur so infrequently or unpredictably in natural settings that resources needed to capture the behavior would be impractical or would otherwise not be observable (Foster & Cone, 1986). For example, non-treatment-seeking alcoholics were exposed to social stress in a laboratory setting and then asked to rate qualities of beer in a mock test designed to assess alcohol consumption following stress exposure (Thomas et al. 2011). Conducting this research in a laboratory setting allowed for testing of hypotheses about stress-potentiated drinking that could not have been done in naturalistic settings. Laboratory settings also may be more convenient and less costly for the researcher.

Naturalistic environments allow observation of behaviors in the setting in which they are most likely to occur, facilitating identification of precise relationships between antecedent and consequent behaviors as well as intervention effects in real-life situations (Tryon, 1998). Utilization of naturalistic settings also allow for observation of individuals, groups, or processes that would not be feasible to recreate in the laboratory. However, naturalistic settings present challenges. Without the protected environment of the laboratory potential external influences on the measures are difficult to control. For example, when collecting data in the home, researchers are confronted with the presence of television or music, household members not involved in the observation (e.g. siblings and pets), and unexpected interruptions from visitors or telephone calls. In clinical settings, observations may be disrupted or missed because of unanticipated changes in health status, diagnostic testing schedules, delivery of therapeutic interventions, and early discharge. The researcher must also be sensitive to the burden of conducting research on families and clinicians. Participant convenience takes priority in scheduling in-home data collection around mealtimes, bedtimes, or other household activities. Observations in a clinical environment have the potential to disrupt caregiving routines and other activities, such as rest times. These problems are not insurmountable, but must be anticipated and addressed when designing observational studies.

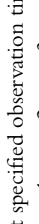
When

Selecting the time frame within which to collect observational data requires a sufficient understanding of the targeted behavior to plan a protocol that will ensure expression of an adequate number of behavioral events (Suen & Ary, 1989). Considerations include time of day when the behavior is most likely to occur or when the behavior is most likely to be elicited (e.g. scheduling challenge tasks for toddlers when rested and recently fed). Within the selected time frame, decisions need to be made about frequency of observations.

Behaviors can be observed either continuously or within a predetermined set of intervals (time sampling). With continuous event recording, the beginning and end of every targeted behavior is recorded. Continuous recording is used to describe sequences of behavior, including latencies (delays) between antecedent and consequent behaviors, and to correlate behavior with concurrently recorded physiologic parameters. Although this type of recording provides the most complete record of the occurrence of a behavior, continuous recordings may not be practical for observations of numerous behaviors or for behaviors that occur at high rates (Hintze & Shapiro, 1995).

Time sampling is used to estimate behavioral frequencies and duration based on a sample of observed behaviors. The observer records the presence or absence of each behavior within a preselected time frame (interval) at a frequency chosen based on the expected occurrence of the target behavior (Hintze & Shapiro, 1995; Skinner et al., 2000). To reduce error in estimation, the shortest possible observation intervals are used that are appropriate to the frequency of the behavior under study (Ary, 1984). Three types of interval recordings may be used: whole interval, partial interval, and momentary time sampling (see Table 6.3; Hintze & Shapiro, 1995).

Table 6.3 Time sampling methods

Type	When	Counting Behavioral Events	Indications	Comments
Whole interval	During entire interval	Behavior counted as occurring only if present during entire interval. Interval 1 2 3 Count 0 1 0 Behavior 	• Useful for assessment of whether behavior is continuous (Hintze & Shapiro, 1995).	1. Underestimates total duration (Powell et al., 1975). 2. More accurate estimates of duration when using short intervals (e.g. 10 seconds; Hintze & Shapiro, 1995).
		Behavior counted as occurring if observed at least once during interval. Interval 1 2 3 Count 1 0 1 Behavior 		
		Behavior 		
Partial interval	During entire interval	Behavior counted as occurring if observed at least once during interval. Interval 1 2 3 Count 1 0 1 Behavior 	• Most useful for non-continuous or sporadic behaviors (Hintze & Shapiro, 1995). • Best method for reliable estimates of frequency (Meany-Daboul et al. 2007).	1. Overestimates duration (Powell et al., 1975) 2. Can be used with longer observation intervals (Hintze & Shapiro, 1995).
Momentary time sampling	At prespecified instant of time, usually at beginning or end of interval	Behavior counted as occurring only if present at specified observation time. Interval 1 2 3 Count * * * Behavior  * = specified observation time	• Best method for reliable estimates of duration (Meany-Daboul et al., 2007; Powell et al., 1977). • Most useful for high rate and/or continuous behaviors or when more than one behavior or participant needs to be observed (Hanley et al., 2007; Skinner et al., 2000).	1. Shorter observation intervals are recommended to enhance accuracy (Foster & Cone, 1986). 2. Useful data can be lost because observations are not recorded between intervals.

Subject	Minute 1						Minute 2					
	10sec	20sec	30sec	40sec	50sec	60sec	10sec	20sec	30sec	40sec	50sec	60sec
1	O R						O R					
2		O R						O R				
3			O R						O R			
4				O R						O R		
5					O R						O R	
6						O R						O R

Figure 6.2 Process of observing multiple participants using momentary time sampling by one observer. White boxes in the illustration indicate observation (box O) and recording (box R). Shaded boxes indicate no observation taking place. Using a predetermined randomly chosen order, the first participant is observed for five seconds and recorded for five seconds. Then the second participant is observed and recorded and the process continues through a total of six participants until one minute is complete. Then recording begins again with the first participant. Using this method each participant has one observation every 60 seconds. The observation frequency can be increased by reducing the number of individuals observed. Blocks of time can be inserted regularly to allow for breaks in observation.

Relationships between estimates and actual occurrence will be greater when behaviors are sampled at 120 seconds or less (Powell et al., 1977).

With whole interval sampling, the behavior is counted only when the behavior occurs throughout the entire interval. While whole interval sampling is not common, the method may be useful when the research interest is on estimates of changes in duration of continuous behaviors, e.g. following an intervention. With partial-interval recording, the observer notes any occurrence of the target behavior during any portion of the observation time, and records it dichotomously as occurrence or non-occurrence, irrespective of the frequency of the behavior within the time interval. This method works well for estimates of less predictable, sporadic behaviors. Momentary time sampling (MTS) occurs when the observer notes whether or not the behavior is occurring only at the instant of the observation. Using MTS, either several different behaviors or a single behavior in several study participants can be observed within the same block of time (see Figure 6.2).

Methods and materials

Various methods and materials are available to record behavioral observations for later analysis. Recent advances in technology have greatly enhanced the types and quantities of behaviors that can be observed and recorded.

Paper-and-pencil

A time-tested, inexpensive, and user-friendly method is a paper-and-pencil based system. These systems involve some form of graph on which the target behaviors are listed in columns and the individual observation times are listed in rows. Checkmarks can be used to mark start and end times of specific behaviors as well as presence or absence of behaviors for which duration is not of interest. Multiple behaviors and behavioral intensities can be observed and coded by placing numeric values in the appropriate cell using a predetermined coding system (see Figure 6.3).

Data Collection Sheet During HRV					
Time	State	Position	Event	Comment	
1 min					
2 min					
3 min					
4 min					
5 min					
6 min					
7 min					
8 min					
9 min					
10 min					
11 min					
12 min					
13 min					
14 min					
15 min					
Codes					
State	1 = Active awake 2 = Quiet awake 3 = Sleep-wake transition 4 = Active sleep 5 = Quiet sleep	Position	1 = Supine 2 = Prone 3 = Side 4 = Semi reclined/ cradeled 5 = Upright	Event	1 = Feeding 2 = Defecating/straining 3 = Position change 4 = Burping/hiccuping 5 = Patient Care 6 = Suctioning

Figure 6.3 Sample data collection sheet with behavioral codes. Behavioral codes are entered each minute and later correlated with physiologic measures.

Comments providing additional information about the observations can be quickly jotted down. Coding in specific increments of time, e.g. 10 seconds, can be accomplished most efficiently by the observer wearing an unobtrusive earpiece to listen to a prerecorded audiotape that simply has the interval number spoken, e.g. interval one . . . , interval two . . . , interval three . . . , etc. (Hintze & Shapiro, 1995). This method allows the observer to focus on observation rather than tracking time intervals.

Electronic recording

With the advent of portable computer systems such as tablets, netbooks, and laptops, efficient recording of behavioral observations may be enhanced. The behavior can be recorded with a mouse click or a touch of a screen or keyboard. The computers can be programmed to automatically time start and end of observed behaviors with simple keystroke combinations. Timers can be set to notify the observer when to observe and when to record. This allows the observer to spend more time on the observation, possibly increasing observational accuracy (Sicotte, 2003). Computers can also be programmed to store the data in a way allowing for direct transfer to spreadsheets or statistical programs, eliminating time-consuming and error-prone data entry necessitated by paper-and-pencil-based systems (Sicotte, 2003).

Audio and video recording

Audio or video recordings are a common method of capturing participant behavior in a wide range of contexts and in response to a number of stimuli. Video recordings are useful for studies in which numerous behaviors or individuals are being observed, or the researcher wishes to

synchronize behavior with continuously observed biophysical or physiological measurements. In laboratory settings, observation suites can be fitted with in-wall cameras and microphones to capture several views with a monitoring room for recording and editing software. When videotaping in naturalistic settings, careful attention to the environment is critical. Participant homes, hospitals, schools, and other institutions where research may be conducted are generally not set up for professional recordings. Key considerations include amount of natural light, location of room lighting, focus point for the camera, and potential for interruptions.

Videotaping behavioral observations can be advantageous. The timing and coding of behaviors can be more precise, and multiple participants and behaviors can be observed and coded. The primary disadvantage to videotaping is that, although videotaping reduces immediate demands on the observer, the time spent in coding is exponentially increased (Suen & Ary, 1989). For example, for the Parent-Child Early Relational Assessment (Clark, 1985, 2010), coding five minutes of video takes an average of one hour. However, 65 items are being coded, which would not be possible in real-time. The benefits of videotaping must clearly outweigh the cost and effort of the additional resources needed.

Audio recordings can be used for more specialized needs. For example, the Language Environment Analysis (LENA) system is an automated vocal analysis system consisting of a digital language processor (for data collection) and speech recognition software (for analysis) developed to measure the language and audio environment of children between the ages of two and 48 months (Dykstra et al., 2013). The child wears a small digital recorder in a pocket for a predetermined duration of time. The recorder distinguishes vocalizations within a four- to six-foot radius (Ford et al., 2008). The system records and processes three main variables: adult word counts, child vocalizations, and child turns (quantifying number of adult-child interactions in which either adult or child responds to the other within five seconds). Because these recordings are able to distinguish words, they may also be used to analyze components of language in specific verbal interactions. The LENA system has been validated in the home (Xu et al., 2009) and in school settings (Dykstra et al., 2013) for measuring influence of adult-child interactions on child language development.

Audio recordings may also be used for non-language-focused observations. In one research program, tiny microphones are placed on the necks of premature infants to hear swallowing and breathing during feeding (Thoyre et al., 2012). Together with concurrent video recording and cardiorespiratory waveforms, these audio recordings of swallowing and breathing enable precise analysis of the infant's ability to coordinate the functions of feeding in preparation for developing interventions to support development of feeding skill.

Actigraphy

Actigraphy is a validated, automated, and cost-effective behavioral observation method for detecting gross motor activity in naturalistic settings. Recordings are analyzed for rest and activity patterns, and data can be used for assessing sleep patterns and sleep disorders (Sadeh, 2011). Actigraphy is most reliable for assessing sleep patterns in healthy individuals with good quality of sleep and in sleep intervention studies. Actigraphy is less reliable for assessing sleep in clinical populations, primarily because of low specificity (i.e., detections of awake time; Sadeh, 2011). Actigraphy recorders are generally contained within a lightweight watch that can be worn on the wrist for children and adults or placed on the lower leg or ankle for infants. Several recorder options are available in which additional information can be obtained, including detection of ambient or colored light, detection and duration of monitor removal, and reminders and response entry for numerically scaled subjective parameters (www.healthcare.philips.com/us_en/homehealth/

sleep/actiwatch/default.wpd). The watches allow recording for between seven and 91 days, depending on desired sampling interval. At least five consecutive days of 24 hour data are needed for a reliable measure of sleep (Sadeh, 2011). Recorders with high sensitivity and at least 60% specificity provide the highest quality results (Sadeh, 2011).

Simultaneous observation of multiple behaviors

Several systems are available for recording and coding multiple types of behavioral observations. Perhaps the most well-known system is the Noldus Behavioral Coding System (Observer), which offers programs for user-defined direct coding of behavioral observations from video, creation of charts and trend lines, and merging of physiological data with behavioral observations (www.noldus.com/human-behavior-research/products/the-observer-xt/). Coding can be precise for frequency and duration of behaviors and can be accomplished in segments, as small as 1/100 of a second.

Ecological momentary assessment

Self-observation, such as ecological momentary assessment (EMA), is a form of direct observation where the observer and the observed are the same person (Cone, 1978). In self-observation, individuals report on their own behaviors in the immediate time and place when prompted by a timer or by the occurrence of a specified event (Nusser, Intille, & Maitra, 2006). With EMA, individuals can report not only on behavioral activities or events, but also associated cognitive, emotional, and physical responses (Shiffman, Stone, & Hufford, 2008). The ability of EMA methods to capture behavior as well as concurrent emotional or cognitive responses enriches the behavioral data and enables examination of effects or outcomes. For example, investigators measured daily emotional dynamics as an outcome measure following a cognitive behavioral therapy intervention in adolescents with major depression using an EMA cell phone protocol (Silk et al., 2011). Participants received randomly timed calls during which they responded to brief interview questions designed to measure current behavior, affect, and social context.

Environmental sensors and bodywide area networks

Sophisticated monitoring systems have been developed to unobtrusively monitor activity and rest in naturalistic settings (Kaye, 2008). These methods do not require the participant to wear a watch or complete a diary, and may be of high value in monitoring health of community-dwelling elderly. One of these systems monitors activity by mounting infrared motion sensors in each room of a research participant's home, allowing calculations of wake and sleep times for individuals, as well as measures of walking speed (Hayes et al., 2008; Hayes et al., 2010). Mobility measured in the home is more accurate than measures done in a clinical setting and may be an early predictor of cognitive impairment (Hayes et al., 2008; Kaye et al., 2012). These motion sensors have also been used to incorporate behavioral feedback into systems to enhance medication adherence through context-based reminders. Hayes and colleagues (2009) placed sensors in homes in each room, on the phone, in the bed, and on the door so the preprogrammed system could determine if and when the medication had been taken. If the medication had not been taken, the system would remind the home dweller at the most opportune time, i.e., when the individual was not in bed, not on the phone, and not away from the house and the time was within a predetermined window (Hayes et al., 2009).

Other types of sophisticated monitoring systems use body area sensor networks to measure activity in conjunction with physiological measurements. These systems can record data from

multiple sensors measuring, for example, blood pressure, heart rate, respiratory rate, pulse oximetry, and limb movement, sending data wirelessly to a network computer that can then be transmitted to health care providers and/or researchers (Ko et al., 2010).

Assessment of data quality

Ensuring the quality of data obtained through direct behavioral observations is critical. Assessment of content validity (Strosahl & Linehan, 1986) and observational accuracy (Cone, 1998) contributes important information about observational data quality.

Content validity

Similar to the methods for testing content validity in surveys, content validity for a behavioral observation is established during the design phase of a study through consultation and review by experts in the field and by potential respondents as well as by analyzing the level of consistency between the identified targeted behaviors and the supporting theory or conceptual framework.

The content validity of the observation itself may be compromised by subject reactivity, i.e., the effect of being observed on the targeted behavior (Foster & Cone, 1986; Suen & Ary, 1989). This threat to content validity can be reduced by sensitizing the participants to observer presence by spending time in the environment prior to beginning data collection and by being as unobtrusive as possible (Suen & Ary, 1989). If the observer's physical presence is highly likely to generate reactivity, consider videotaping. Reactivity to videotaping may be reduced by placing the video camera in an unobtrusive location or behind a two-way mirror. In addition, using middle rather than initial recording segments for coding allows the participant time to habituate to the camera before the behavior is measured. Haidet and colleagues (2009) provide an excellent review of methods to improve reliability of video recorded behavioral observations.

Observational accuracy

As with any research measure, direct observational measures are only as valid (i.e., accurately measuring the concept of interest) as they are reliable (i.e., consistent). To ensure observational accuracy, thoroughly train observers to objectively record target behaviors and proactively address threats to observational accuracy.

Observer training

Observers need to be trained in a methodical manner. Training videos provide an efficient resource for teaching and testing specific behavioral coding schemes. To develop training videos, record samples of the behavior to be observed and code them based on study-developed criteria. Then review and discuss the training videos with the observers-in-training to assist them in identifying and applying coding criteria (Tryon, 1998). Next, ask the observers-in-training to independently observe additional training video recordings, repeating with new recordings until acceptable interrater reliability has been reached, e.g., *kappa* values greater than .60 (Landis & Koch, 1977). At this point, the observer-in-training can perform observations independently with regular, random reliability assessments by a second trained observer to ensure consistent recording of observations.

Threats to observational accuracy

Threats to observational accuracy include observer drift and observer bias (Foster & Cone, 1986). *Observer drift* refers to trained observers gradually altering their ratings of the same behavior over time (Tryon, 1998). Even highly trained observers are subject to drift, especially when the observations are repetitive or the observation sessions are lengthy (Suen & Ary, 1989). *Observer bias* occurs when observer ratings are distorted by observer expectations (Foster & Cone, 1986; Suen & Ary, 1989) or by observer fatigue (Tryon, 1998). *Observer expectations* refers to expectations the observer may theoretically have about the behavior based on the observer's personal knowledge of study hypotheses or perceptions the observer has about the population being studied. *Observer fatigue* may occur when the duration of the observation period is lengthy or when the behaviors targeted for observation are too numerous (Tryon, 1998). These potential contributors to observational accuracy can be reduced using multiple strategies listed in Table 6.4.

Interrater agreement and reliability

Observations require a component of human judgment, resulting in disagreements between raters in kind (e.g., normal vs abnormal gait), grade (e.g., mild, moderate, or severe asthma), or magnitude (e.g., degree of maternal sensitivity in interaction with her newborn). Rater disagreement can be reduced, but not completely eliminated, by attention to the training and prevention strategies discussed above. Evaluation of disagreement between raters is an essential component of observational research design using measures of interrater agreement (IRA) and interrater reliability (IRR). Agreement refers to the extent to which raters score items in exactly the same way; reliability refers to how well the raters are able to consistently distinguish between various items (Gisev, Bell, & Chen, 2013).

In assessing IRA and IRR, the possibility of chance agreement must be accounted for. The various kappa statistics correct for chance agreement and can be used for both IRA and IRR

Table 6.4 Threats to observational accuracy

<i>Threat</i>	<i>Definition</i>	<i>Strategies to Reduce Threats</i>
Drift	Gradual alteration of ratings of the same behavior over time (Tryon, 1998)	Clear descriptions of molecular behaviors Extensive initial training to reliability and regular recalibration of observational skills Random reliability checks of at least 20% of observations distributed throughout the study (Snyder et al., 2006)
Expectation Bias	Expectations the observer may theoretically have about the behavior based on knowledge of study hypotheses or perceptions the observer has about the population being studied that may affect observer ratings (Foster & Cone, 1986; Suen & Ary, 1989)	Clear descriptions of molecular behaviors Thorough and on-going observer training When possible, keep observers blinded to study hypotheses (Suen & Ary, 1989)
Fatigue	Occurs when the duration of the observation or coding periods are lengthy or when the target behaviors are too numerous (Tryon, 1998)	Select a manageable number of observable target behaviors Provide for regular breaks in observation sessions and while coding audiovisual materials

indices (Gisev et al., 2013). For nominal outcomes, Cohen's kappa is used; weighted kappa or Kendall's coefficient of concordance are used on ordinal measurements (Gisev et al., 2013). Interrater reliability studies using interval or ratio measurements typically involve n targets (i.e., specific focus of the observation) and k raters. Each target may be evaluated by: (a) a different set of k judges randomly selected from a population of judges; (b) a random set of k judges that rates each target; or (c) the same k judges, who are the only judges of interest (Shrout & Fleiss, 1979, p. 421). One- or two-way mixed or random effects analysis of variance models are specified to decompose variance into portions due to targets (true variance) and due to judges (regarded as extraneous in many situations), considering whether absolute or consistent ratings and single or average ratings are of interest (McGraw & Wong, 1996; Shrout & Fleiss, 1979).

Conclusion

Direct behavioral observation is a valuable tool for measuring behavior in research. With careful attention to developing and defining the behaviors to observe, the context in which to observe, and the methods for assuring data quality, behavioral observation can be applied to any population and in most settings in which research could be conducted. Technological advances to assist in observing behavior create opportunities to develop innovative methods to efficiently and accurately describe health behaviors and to develop and test interventions to enhance the health of individuals, families, and communities.

References

- Ary, D. (1984). Mathematical explanation of error in duration recording using partial interval, whole interval, and momentary time sampling. *Behavioral Assessment*, 6, 221–228.
- Campbell, M. L., Templin, T., & Walch, J. (2010). A respiratory distress observation scale for patients unable to self-report dyspnea. *Journal of Palliative Medicine*, 13, 285–290.
- Clark, T. (1985, 2010). *The parent-child early relational assessment: Instrument and manual*. Madison: Department of Psychiatry, University of Wisconsin Medical School.
- Clayman, M. L., Makoul, G., Harper, M. M., Koby, D. G., & Williams, A. R. (2012). Development of a shared decision making coding system for analysis of patient-healthcare provider encounters. *Patient Education and Counseling*, 88, 367–372.
- Cone, J. D. (1978). The behavioral assessment grid (BAG): A conceptual framework and a taxonomy. *Behavior Therapy* 9, 882–888.
- Cone, J. D. (1998). Psychometric considerations: Concepts, contents, and methods. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston, MA: Allyn and Bacon.
- de Lemus, S., Spears, R., & Moya, M. (2012). The power of a smile to move you: Complementary submissiveness in women's posture as a function of gender salience and facial expression. *Personality and Social Psychology Bulletin*, 38, 1480–1494.
- Dykstra, J., Sabatos-DeVito, M. G., Irvin, D. W., Boyd, B. A., Hume, K. A., & Odom, S. L. (2013). Using the language environment analysis (LENA) system in preschool classrooms with children with autism spectrum disorders. *Autism*, 17, 582–594.
- Ford, M., Baer, C. T., Xu, D., Yapanel, U., & Gray, S. (2008). The LENA™ language environment analysis system: Audio specifications of the DLP-0121 (LENA Foundation Technical Report LTR-03–2). Retrieved from www.lenafoundation.org/TechReport.aspx/Audio_Specifications/LTR-03-2
- Foster, S. L., & Cone, J. D. (1986). Design and use of direct observation procedures. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (2nd ed.) (pp. 253–324). New York, NY: Wiley.
- Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9, 330–338.
- Haidet, K. K., Tate, J., Divirgilio-Thomas, D., Kolanowski, A., & Happ, M. B. (2009). Methods to improve reliability of video-recorded behavioral data. *Research in Nursing and Health*, 32, 465–474.
- Hanley, G. P., Cammilleri, A. P., Tiger, J. H., & Ingvarsson, E. T. (2007). A method for describing preschoolers' activity preferences. *Journal of Applied Behavior Analysis*, 40, 603–618.

- Harrison, T.M. (2009). Effect of maternal behavior on regulation during feeding in healthy infants and infants with transposition. *Journal of Obstetric, Gynecologic, and Neonatal Nursing*, 38, 504–513.
- Hayes, T. L., Abendroth, F., Adami, A., Pavel, M., Zitzelberger, T. A., & Kaye, J. A. (2008). Unobtrusive assessment of activity patterns associated with mild cognitive impairment. *Alzheimer's & Dementia*, 4, 395–405.
- Hayes, T. L., Cobbinah, K., Dishongh, T., Kaye, J. A., Kimmel, J., Labhard, M., . . . Vergun, S. (2009). A study of medication-taking and unobtrusive, intelligent reminding. *Telemedicine and e-Health*, 15, 770–776.
- Hayes, T. L., Riley, T., Pavel, M., & Kaye, J. A. (2010, August). *Estimation of rest-activity patterns using motion sensors*. IEEE, 32nd Annual International Conference of the IEEE EMBS, Buenos Aires, Argentina.
- Heathcock, J. C., Bhat, A. N., Lobo, M. A., Galloway, J. C. (2004). The performance of infants born preterm and full-term in the mobile paradigm: Learning and memory. *Physical Therapy*, 84, 808–821.
- Held, J., & Manser, T. (2005). A PDA system for on-line recording and analysis of concurrent events in complex behavioral processes. *Behavior Research Methods*, 37, 155–164.
- Hintze, J. M., & Shapiro, E. S. (1995). Best practices in the systematic observation of classroom behavior. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (3rd ed.) (pp. 651–660). Washington, DC: National Association of School Psychologists.
- Hintze, J. M., Stoner, G., & Bull, M. H. (2000). Analogue assessment: Emotional/behavioral problems. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Conducting school-based assessments of child and adolescent behavior* (pp. 55–77). New York, NY: Guilford.
- Jovanov, E., Milenkovic, A., Otto, C., DeGroen, P., Johnson, B., Warren, S., & Taibi, G. (2005, September). *A WBAN system for ambulatory monitoring of physical activity and health status: Applications and challenges*. Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China.
- June, K. M., Norton, K. J., Rees, V. W., & O'Connor, R. J. (2012). Influence of measurement setting and home smoking policy on smoking topography. *Addictive Behaviors*, 37, 42–46.
- Kaye, J. (2008). Home based technologies: A new paradigm for conducting dementia prevention trials. *Alzheimer's & Dementia*, 4 (1 Suppl. 1), S60–S66.
- Kaye, J., Mattek, N., Dodge, H., Buracchio, T., Austin, D., Hagler, S., . . . Hayes, T. (2012). One walk a year to 1000 within a year: Continuous in-home unobtrusive gait assessment of older adults. *Gait & Posture*, 35, 197–202.
- Kendler, K. S., Halberstadt, L. J., Butera, F., Myers, J., Bouchard, T., & Ekman, P. (2008). The similarity of facial expressions in response to emotion-inducing films in reared-apart twins. *Psychological Medicine*, 38, 1475–1483.
- Ko, J., Lu, C., Srivastava, M. B., Stankovic, J. A., Terzis, A., & Welsh, M. (2010). Wireless sensor networks for health care. *Proceedings of the IEEE*, 98, 1947–1960.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- Lavelle, M., Healey, P. G. T., & McCabe, R. (2013). Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia Bulletin*, 39, 1150–1158.
- Loman, M. M., Johnson, A. E., Westerlund, A., Pollak, S. D., Nelson, C. A., & Gunnar, M. R. (2013). The effect of early deprivation on executive attention in middle childhood. *Journal of Child Psychology and Psychiatry*, 54, 37–45.
- McGrath, J. E., & Tschan, F. (2004). *Temporal matters in social psychology*. Washington, DC: American Psychological Society.
- McGraw, K. O., & Wong, S. P. (1996) Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–16.
- Meany-Daboul, M. G., Roscoe, E. M., Bourret, J. C., & Ahearn, W. H. (2007). A comparison of momentary time sampling and partial-interval recording for evaluating functional relations. *Journal of Applied Behavior Analysis*, 40, 501–514.
- Meeks, S., & Looney, S. W. (2011). Depressed nursing home residents' activity participation and affect as a function of staff engagement. *Behavior Therapy*, 42, 22–29.
- Merrell, K. W. (2000). Informant report: Rating scale measures. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Conducting school-based assessments of child and adolescent behavior* (pp. 203–234). New York, NY: Guilford.
- Mesman, J., Alink, L. R. A., van Zeijl, J., Stolk, M. N., Bakermans-Kranenburg, M. J., van IJzendoorn, M. H., . . . Koot, H. M. (2008). Observation of early childhood physical aggression: A psychometric study of the system for coding early physical aggression, *Aggressive Behavior*, 34, 539–552.
- Mitchell, M., Piazza-Waggoner, C., Modi, A., & Janicke, D. (2009). Examining short-term stability of the mealtime interaction coding system (MICS). *Journal of Pediatric Psychology*, 34, 63–68.

- Nettle, D. (2011). Large differences in publicly visible health behaviors across two neighborhoods in the same city. *PLoS ONE*, 6(6), e21051.
- Nusser, S. M., Intille, S. S., & Maitra, R. (2006). Emerging technologies and next-generation intensive longitudinal data collection. In T. A. Walls & J. L. Schafer (Eds.), *Models for intensive longitudinal data* (pp. 254–278). Oxford: Oxford University Press.
- Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis*, 8, 463–469.
- Powell, J., Martindale, B., Kulp, S., Martindale, A., & Bauman, R. (1977). Taking a closer look: Time sampling and measurement error. *Journal of Applied Behavior Analysis*, 10, 325–332.
- Reed, J. A., Hooker, S. P., Muthukrishnan, S., & Hutto, B. (2011). User demographics and physical activity behaviors on a newly constructed urban rail/trail conversion. *Journal of Physical Activity and Health*, 8, 534–542.
- Robbins, M. L., Mehl, M. R., Holleran, S. E., & Kasle, S. (2011). Naturalistic observed sighing and depression in rheumatoid arthritis patients: A preliminary study. *Health Psychology*, 30, 129–133.
- Roberts, L. J., Luke, D. A., Rappaport, J., Seidman, E., Toro, P. A., Reischl, T. M. (1991). Charting uncharted terrain: A behavioral observation system for mutual help groups. *American Journal of Community Psychology*, 19, 715–737.
- Sadeh, A. (2011). The role and validity of actigraphy in sleep medicine: An update. *Sleep Medicine Reviews*, 15, 259–267.
- Shapiro, E. S., & Kratochwill, T. R. (2000). Introduction: Conducting a multidimensional behavioral assessment. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Conducting school-based assessments of child and adolescent behavior* (pp. 1–20). New York, NY: Guilford.
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4, 1–32.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Validity*, 86, 420–428.
- Sicotte, C. (2003). *Computerized behavioral observation: Implications for improving accuracy and reducing bias* (Doctoral dissertation). Retrieved from ProQuest Information and Learning (3103725).
- Silk, J. S., Forbes, E. E., Whalen, D. J., Jakubcak, J. L., Thompson, W. K., Ryan, N. D., . . . Dahl, R. E. (2011). Daily emotional dynamics in depressed youth: A cell phone ecological momentary assessment study. *Journal of Experimental Child Psychology*, 110, 241–257.
- Skinner, C. H., Rhymer, K. N., & McDaniel, E. C. (2000). Naturalistic direct observation in educational settings. In E. S. Shapiro & T. R. Kratochwill (Eds.) *Conducting school-based assessments of child and adolescent behavior* (pp. 21–54). New York, NY: Guilford.
- Snyder, J., Reid, J., Stoolmiller, M., Howe, G., Brown, H., Dagne, G., & Cross, W. (2006). The role of behavior observation in measurement systems for randomized prevention trials. *Prevention Science*, 7, 43–56.
- Strosahl, K. D., & Linehan, M. M. (1986). Basic issues in behavioral assessment. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (2nd ed., pp. 12–46). New York, NY: Wiley.
- Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Lawrence Erlbaum.
- Thomas, S. E., Bacon, A. K., Randall, P. K., Brady, K. T., & See, R. E. (2011). An acute psychosocial stressor increases drinking in non-treatment-seeking alcoholics. *Psychopharmacology*, 218, 19–28.
- Thoyre, S. M., Holditch-Davis, D., Schwartz, T. A., Roman, C. R. M., & Nix, W. (2012). Coregulated approach to feeding preterm infants with lung disease. *Nursing Research*, 61, 242–251.
- Tryon, W. W. (1998). Behavioral observation. In A. S. Bellack & M. Hersen (Eds.) *Behavioral assessment: A practical handbook* (4th ed., pp. 79–103). Boston, MA: Allyn and Bacon.
- van Oorsouw, W. M. W. J., Embregts, P. J. C. M., & Sohier, J. (2011). Verbal and non-verbal emotional behavior of staff: A first attempt in the development of an observation instrument. *Research in Developmental Disabilities*, 32, 2408–2414.
- Vázquez, M., Pardavila, M., Lucia, M., Aguado, Y., Margall, M. A., Asiain, M. C. (2011). Pain assessment in turning procedures for patients with invasive mechanical ventilation. *Nursing in Critical Care*, 16, 178–185.
- Xu, D., Yapanel, U., & Gray, S. (2009). Reliability of the LENA language environment analysis system in young children's natural language home environment (LENA Foundation Technical Report LTR-05-2). Retrieved from www.lenafoundation.org/TechReport.aspx/Reliability/LTR-05-2

7

BIOPHYSICAL MEASUREMENT

Ulf G. Bronas and Diane Treat-Jacobson

Human biological processes are rooted in biophysics. Biophysical measures used in clinical research are based on principles related to chemical energy, mechanical (kinetic) energy, temperature or heat, sound, light, electricity, electromagnetics, pressure, and flow.

Biophysical principles and methods can be traced back to the 5th-century BCE Greek scientist Heraclitus (Glaser, 2010). Today, biophysicists study biomolecular systems and develop methods, tools, and computational methodologies using the principles of physics to study complex biological systems. Nursing scientists use biophysical principles and methods to fully investigate health-illness processes and effects of clinical interventions.

The purpose of this chapter is to provide an overview of key biophysical principles used in nursing research. Content is selective, emphasizing topics relevant to current problems in nursing science. Principles are introduced and explained; measurement methods are described; and examples of use in assessment, diagnosis, and intervention in research are provided.

Biophysics used in health care and research in the health sciences is increasingly complex. Recently published resources contain information with varying topical coverage and difficulty levels. Davidovits (2013) summarizes the application of topics in physics to biology and medicine. Dillon (2012) is an excellent and easy-to-read book that explains biophysics in the context of physiology. Cotterill (2008) provides an outstanding introduction to the field of biophysics on a basic science level from atoms to organisms. Glaser (2010) gives an in-depth mathematical review of molecular and supramolecular structure and function to complex biological systems. Sharma (2011) provides an easy-to-read practical application of biophysics to nursing and nursing care.

General principles of energy

Energy in various forms is central to biophysics. Chemical energy enables cells to grow, maintain a suitable environment, and maintain life processes. Chemical energy used for active transport enables movement of ions, molecules, and other particles across cell membranes, and is crucial in creating concentration gradients. Mechanical energy enables movement either at the cellular level or at the whole body level (which also involves movement of filaments or fibers). Biological systems store potential energy in concentration gradients and chemical bonds, which are transformed to kinetic energy to produce chemical, transport, or mechanical work. Energy external

to the body, like light, sound, and electricity, is used for assessment, diagnosis, and in interventions to promote health and prevent or alleviate illness conditions.

Energy transfer

The laws of thermodynamics govern energy transfer in biological systems. The first law of thermodynamics states that, in a closed system, the total amount of energy is constant. Thus, energy cannot be created or destroyed; it can only change forms. Humans obtain energy from outside the system, primarily as glucose (3.8 kcal/gram), other carbohydrates (4.1 kcal/gram), proteins (4.1 kcal/gram), or fats (9.3 kcal/gram); humans lose energy to the surroundings (primarily as heat) (Davidovits, 2013). The change in energy ΔU is commonly depicted as

$$\Delta U = q - W, \quad (7.1)$$

where q is heat put into the system and W is work done. Work is related to heat; both involve transfer of energy from one system to another, but heat transfers energy only by thermal interaction via conduction, convection, or radiation. Heat is more readily measured than work.

Enthalpy, entropy, and Gibbs free energy

Enthalpy is an expression of system energy changes through chemical or biological processes. It is the sum of the internal energy of a system plus its pressure multiplied by its volume; this is the classic thermodynamic property of a system, or a way to divide energy into different components (i.e., energy associated with pressure change and energy associated with heat flowing into the system). Most physiological processes occur at constant pressure, so that enthalpy changes equal the heat flowing into the system (Cotterill, 2008, Chapter 4; Glaser, 2010, Chapter 2), shown as

$$q = \Delta H, \quad (7.2)$$

where q is heat flowing into the system and ΔH is change in enthalpy.

The second law of thermodynamics states that natural and spontaneous processes move from a state of order to a state of disorder (i.e., entropy). Entropy is energy lost due to dissipation (not used for work), whereas Gibbs free energy is energy available to actually do work such that

$$G = H - T - S, \quad (7.3)$$

where G is Gibbs free energy, H is enthalpy, T is absolute temperature, and S is entropy. Thus, energy free to do work (Gibbs free energy) is energy left from enthalpy after entropy has been removed. This equation shows that if a system goes from a higher energy-state to a lower energy-state, energy is released, and thus available for work (Chodera & Mobley, 2013; Cottrill, 2008; Dillon, 2012, Chapter 1; Heermann, 2011).

Chemical energy

Chemical energy is potential energy stored in the bonds of atoms and molecules; it allows a substance to be transformed through a chemical reaction (Freire, 2008). Chemical energy is a basic component of physiological functions and provides the capacity to do work. Human bodies metabolize food into usable forms of energy to perform basic physiological functions, maintain body temperature, and enable molecular and whole body movements. Most unused energy is

stored; a small percentage is excreted as waste. The use of energy to sustain basic body functions is reflected by the resting energy requirements of the body, which are approximately 92 cal/kg (Sharma, 2011, Chapter 4).

All living organisms need energy to sustain life and life processes. During glycolysis, glucose is converted to pyruvate. Energy released during this reaction is used to produce adenosine triphosphate (ATP), a high-energy compound. When high-energy bonds are broken in the conversion of ATP into adenosine diphosphate (ADP), energy is released. Free energy released during conversion of ATP to ADP can be used immediately for work or can be stored as sugars or fat for later use, ATP is used as an energy source to drive everything from intracellular ionic pumps to whole body motion (Cotterill, 2008, Chapter 4; Dillon, 2012, Chapter 9).

Measurement: resting metabolic rate

Resting or basal metabolic rate (BMR) is caloric expenditure at complete rest (Davidovits, 2013). The current gold standard for measuring BMR is flow-through indirect calorimetry, which requires a laboratory setting and expensive equipment. Consequently, BMR is often estimated using a predictive equation based on weight and height. Using predictive equations is relatively easy, but accuracy varies with the patient population, precision of equipment used to measure height and weight, and equation used. Predictive equations may be more appropriate for use in healthy individuals than for those in ill health (Madden, Parker, & Amirabdollahian, 2013). The Mifflin-St. Jeor equations (Mifflin, St. Jeor, Hill, Scott, Daugherty, & Koh, 1990) for males is

$$\text{RMR (males)} = (9.99 \times \text{weight [kg]}) + (6.25 \times \text{height [cm]}) - (4.92 \times \text{age [years]}) + 5$$

and for females is

$$\text{RMR (females)} = (9.99 \times \text{actual weight}) + (6.25 \times \text{height}) - (4.92 \times \text{age}) - 161.$$

Applications: chemical energy

Food is converted to stored energy. As metabolic demand is increased, stored energy is accessed. These processes are significantly altered during varied states of health and illness. Understanding changes in nutritional requirements and metabolism during different illness states has significant importance for health status. For example, nutritional balance is related to clinical outcomes and length of stay in the ICU (Singer et al., 2014) and across the continuum of care (de van der Schueren et al., 2014).

Understanding the nutritional needs of infants and children and the nutritional components contributing to optimal growth and development and normal weight is essential. Dietary interventions for at-risk mothers before and during pregnancy have improved birth weight and reduced incidence of low birth weight (Gresham, Byles, Bisquera, & Hure, 2014). Children hospitalized with nonserious illness are known to experience significant nutritional deterioration (Pacheco-Acosta et al., 2014). Children between ages one to three have increased nutritional requirements relative to adults at a critical time when the diet rapidly changes and new nutritional patterns are established (Alles, Eussen, & van der Beek, 2014). Suboptimal nutrition early in life is associated with future risk of chronic disease (Lillicrop & Burdge, 2011).

Obesity is a preventable worldwide epidemic associated with diabetes, cardiovascular disease, musculoskeletal disorders, and some cancers (World Health Organization, 2014). Understanding principles of energy metabolism is essential to design of interventions for obesity prevention

and weight loss, many of which involve ongoing modification of dietary intake and exercise (National Heart, Lung, and Blood Institute, 2013). In a meta-analysis of 24 randomized controlled trials involving 1,063 participants, an isocalorically prescribed high protein diet provided modest benefit for reductions in body weight compared with an energy-restricted standard-protein diet (Wycherley, Moran, Clifton, Noakes & Brinkworth, 2012).

Mechanical energy

Mechanical energy is a combination of potential and kinetic energy (*KE*). Objects in motion have kinetic energy. Greater mass (*m*) and speed (velocity; *v*) generate more kinetic energy (Davidovits, 2013):

$$KE = \frac{1}{2} m \times v^2. \quad (7.4)$$

Heat and work are not properties of biological systems such as persons or animals, but are properties of processes that transfer energy. In general, the amount of heat or work in an object cannot be measured; instead, amount of energy transferred among objects in certain ways during the occurrence of a given process is measureable (Holtzclaw, 2001).

Bodies move, and have kinetic energy, when they are physically active. Physical activity is “any bodily movement produced by skeletal muscles that results in a substantial increase over resting energy expenditure” (American College of Sports Medicine, 2010, p. 2). Definitions of exercise vary, but all incorporate physical activity. Exercise is commonly regarded as planned, recurring subset of physical activity that results in physical fitness (cardiorespiratory fitness, muscle strength, body composition, and flexibility) (Thompson et al., 2003; Treat-Jacobson, Bronas, & Salisbury, 2013). The efficacy with which physical activity improves health outcomes is well-documented (American College of Sports Medicine [ACSM], 2010).

Measurement: exercise and physical activity outcomes

Physical activity and functional outcomes affected by physical activity are central to nursing science. Mode (walking, running; or, aerobic vs anaerobic, weight bearing or nonweight bearing), frequency (sessions, days), duration, and intensity (physical effort required to perform an activity) are all relevant to measurement of physical activity. Aerobic exercise is measured in metabolic equivalents (METs) (ACSM, 2010).

Physical activity can be measured objectively and subjectively. Pedometers and accelerometers (e.g., Fitbit Flex and Garmin Vivofit) are reliable and valid for activity monitoring and determining energy expenditure when walking or jogging without an incline (Adam Noah, Spierer, Gu, & Bronner, 2013); their use can facilitate validation of subjective physical activity reports. In addition, wearable devices can provide feedback on cumulative daily physical activity and can also be used as an intervention to motivate participants to reach daily goals (e.g., Harris et al., 2013).

Numerous validated self-report questionnaires are available. The 7-Day Physical Activity Recall Questionnaire (PAR; Sallis et al., 1985) is a semi-structured interview that assesses amount of time spent sleeping and engaged in moderate, hard, and very hard activity. Scores on the 7-Day PAR were valid in multiple populations and moderately correlated with accelerometer data in older adults (Sloane et al., 2009). Sometimes, it is more important to measure flexibility, balance, or strength than aerobic activity. This is particularly true of studies involving older participants, such as studies testing interventions to prevent falls or maintain independent living (Chen &

Chou, 2013). The Short Physical Performance Battery (SPPB) (Gurlanik, 2012) measures lower extremity physical functioning objectively (getting up from a chair, balance, gait speed) in older adults. The 6-Minute Walk Test is a useful functional test performed in a corridor without specialized equipment (American Thoracic Society, 2002).

Application: physical activity and exercise interventions

Exercise interventions are widely used in nursing research to address numerous health problems in populations of all ages (Novak, 2011; Ruppert, Conn, Chase, & Phillips, 2014; Keteyian, Fleg, Brawner, & Piña, 2010; Pinto & Floyd, 2007). Duration of exercise intervention, the duration of each session, the mode (type), and intensity of exercise (ACSM, 2010) combined result in the “dose” of exercise.

Thermal energy (temperature; heat)

Temperature influences reaction rates of various chemical and biological processes (e.g., diffusion, cell division, enzymatic reactions) (Bisson & Younker, 2006; Davidovits, 2013; Dillon, 2012, Chapters 1–4). Random motion of molecules produces heat, which is difficult to measure. Instead, relative warmth of a system can be assessed by measuring the effects of heat on the movement of molecules. Temperature reflects the heat content of a substance and is the measure of the mean kinetic energy of molecules within a given system. Body temperature is maintained by continuous interaction with the environment, which occurs through radiation, conduction, convection, or evaporation (Holtzclaw, 2001).

Measurement: temperature

Thermometers are temperature-sensing instruments (Holtzclaw, 1998) calibrated on Celsius, Kelvin, and Fahrenheit scales. The Kelvin scale is used in physics; the Celsius scale is used in nursing research (Table 7.1).

Temperature can be measured at numerous sites in the human body: the pulmonary artery, oral cavity, rectum, tympanic membrane and temporal artery, and forehead. Each site has advantages and disadvantages in various situations. There can be a trade-off between accuracy and

Table 7.1 Temperature scales

<i>Scales</i>	<i>Indicator</i>	<i>Characteristics</i>	<i>Transformations</i>
Celsius	°C	• 0°: melting point of ice • 37°: average normal human body temperature • 100°: steaming point of water	$^{\circ}\text{C} = (5/9)^{\circ}\text{F} - 32$ $0^{\circ}\text{C} = 32^{\circ}\text{F}$ $100^{\circ}\text{C} = 212^{\circ}\text{F}$ $-40^{\circ}\text{C} = -40^{\circ}\text{F}$
Fahrenheit	°F	• 32°: water freezes to ice^a • 212°: water boils^a • Not used for scientific work	$^{\circ}\text{F} = (9/5)^{\circ}\text{C} + 32$
Kelvin	K	• Absolute thermodynamic temperature scale • All thermal motion ceases at 0 K	$0\text{ K} = -273.15^{\circ}\text{C}$ $0\text{ K} = -459.67^{\circ}\text{F}$

^a At sea level and standard atmospheric pressure.

invasiveness. Temperature measurements taken from a pulmonary artery catheter, distal esophagus, or nasopharynx more accurately reflect core body temperature and are less susceptible to measurement error, but are highly invasive and not appropriate outside of perioperative or critical care settings (Hooper & Andrews, 2006; Pusnik & Drnovsek, 2005). Oral thermometry is relatively safe and easy to perform and closely estimates pulmonary artery temperature, even in acutely ill patients, if performed correctly (Hooper & Andrews, 2006). Oxygen therapy, mouth breathing, and oral lesions can influence the accuracy of oral temperature measurement (Holtzclaw, 1998). Infrared tympanic thermometry is increasingly common in the acute care setting, but showed low sensitivity for diagnosing fever in children compared to rectal temperature measurement (Dodd, Lancaster, Craig, Smyth, & Williamson, 2006), and was the least accurate thermometry device in perioperative patients (Dodd et al., 2006).

Accurate, precise assessment and optimization of temperature is critical in many situations. Perioperative hypothermia is related to perioperative morbidity in adults, and prevention of perioperative hypothermia has improved perioperative complications (Forbes et al., 2009). Critically ill and premature neonates lose heat easily. Thus, their environments must be carefully temperature-controlled to maximize the amount of energy they have for growth and development and to prevent disease (Seden, 2011; Smith, Alcock, & Usher, 2013).

Interventions: temperature

Manipulation of temperature is an increasingly common therapeutic intervention. Therapeutic hypothermia following cardiac arrest can have desirable effects on temperature-dependent cellular reactions; therapeutic hypothermia can decrease metabolic and oxygen requirements of tissue (Alkadri, Peters, Katz, & White, 2013), and thus decrease cardiac muscle damage. Direct manipulation of the environment to control temperature is routinely done during exercise testing to increase comfort of participants and standardize testing conditions (ACSM, 2010).

Hyperthermia is also used as a therapeutic intervention. In cancer therapy, hyperthermia may be an effective adjunct to chemotherapy in patients with soft tissue sarcoma (Issels et al., 2010).

Application of targeted neck cooling resulted in significant decrease in pain in patients with migraine (Sprouse-Blum, Gabriel, Brown, & Yee, 2013). Both cooling and heating have been used to treat symptoms following musculoskeletal injury. Typically, ice has been indicated for acute injuries following the theory that application of cold will decrease inflammation, pain, and edema. Heat is more often recommended for chronic injuries. Heat in conjunction with exercise therapies has been shown to reduce pain and improve function, although the evidence is not strong (French, Cameron Walker, Reggars, & Esterman, 2006).

Sound

Sound is a mechanical wave resulting from particle vibrations. Sound waves are transmitted through a medium (like air); the speed of transmission depends on the medium through which the sound waves travel. Sound waves are translated into air vibrations when they reach the human ear, and are interpreted by the brain as sound or noise (Davidovits, 2013).

Measurement: sound

Sound wave frequency is measured in Hertz (Hz; number of cycles per second). The human ear can detect sound frequencies from 20–20,000 Hz and is most sensitive from 200–4,000 Hz (Davidovits, 2013). The infrasonic range is < 20 Hz; ultrasonic range is > 20,000 Hz (Cobbold,

2006; Webb, 2002, Chapter 3). Perception of pitch increases as the frequency of sound waves increases. Sound wave intensity (noise) is measured in decibels (dB). Noise intensity ranges from 0 dB (hearing threshold) to 120 dB (pain threshold) (Davidovits, 2013). Sound level is measured using meters that incorporate a nondirectional condenser microphone, attenuator, amplifier, and weighting and filter networks (Hughes, 2007).

The U.S. Environmental Protection Agency has recommended an average 24-hour exposure limit of 55 A-weighted decibels (dBA) to protect from all health-related adverse effects and a second exposure limit of 70 dBA to prevent hearing loss (Hammer, Swinburn, & Neitzel, 2014). The American Academy of Pediatrics recommends a noise level of no more than 45 dB in the neonatal intensive care unit (NICU) (Allen, 2013; American Academy of Pediatrics, 1997).

Applications: sound

Sound is used during physical assessment to evaluate the status of internal systems. The most common method of assessment is through auscultation, usually using a stethoscope (Davidovits, 2013; Table 7.2), but other instrumentation can be used (e.g., sound recording for objective measurement of cough; Smith & Woodcock, 2008). Auscultation can be used to determine whether bodily sounds are within or outside standards that represent healthy functioning, including blood pressure, heart sounds, arterial or venous bruits, lung sounds, and bowel sounds.

Numerous nursing interventions utilize sound as a therapy, or employ methods to reduce environmental sound (noise) to protect individuals from excessive environmental sound for therapeutic benefit. Examples of sound-based interventions are listed in Table 7.3.

Ultrasound

Ultrasound is produced by electrically pulsed current resulting in mechanical vibration that produces very high frequency sound waves. Electrical current sent to a transducer head containing a piezoelectric crystal causes the crystal to expand and contract. This vibration is converted from electrical energy to acoustic energy (sound waves). The sound wave must be transmitted through a conduction medium (e.g., ultrasound gel). The higher-frequency sound waves are absorbed in more superficial tissues whereas lower-frequency sound waves penetrate more deeply. The sound wave is attenuated as it travels through the tissue due to absorbance, reflection, or dispersion. Tissues with high water content have a fairly low absorption rate, whereas denser tissues (with high protein content) have a higher absorption rate. As sound waves pass through tissues, part of the energy is scattered (reflected or refracted). The extent of reflection is determined by the acoustic impedance between the different tissues. The greater the difference in acoustic impedance between tissues, the greater the amount of energy reflected back to the transducer. An image is produced by measuring the rate of reflection and the differences in acoustic impedance between tissues (Bouchet et al., 2001; Metter, Beutel, & Kundel, 2000, Chapter 7; Webb, 2002, Chapter 3).

Real-time mode ultrasonography at 30 frames of compound cross-sectional B-mode images per second is used to create three-dimensional images and depict blood flow (Bouchet et al., 2001; Metter et al., 2000, Chapter 7). Recent advances in the size and portability of ultrasound devices permit use at point-of-care, which can facilitate diagnostic efficiency and accuracy (Gillman & Kirkpatrick, 2012).

Duplex ultrasound permits noninvasive measurement of flow in blood vessels, and is useful for diagnosis of arterial or venous stenosis, deep venous thrombosis, and venous insufficiency (Figure 7.1). Ultrasound can be applied for therapeutic effects. Continuous ultrasound creates heat, which can be used to relieve pain and promote healing of injuries (e.g., Lesniak et al., 2014;

Table 7.2 Bodily sounds assessed through auscultation with a stethoscope

<i>Sound</i>	<i>Definition</i>	<i>Implications</i>
Systolic BP/1st Korotkoff sound	- Maximal pressure during left ventricular contraction - Point at which pressure inside the artery exceeds the pressure exerted by the blood pressure cuff - Sounds created by turbulent flow in the blood vessel caused by partial constriction of the artery	- BP outside normal limits can significantly impact CV health - Hypertension is related to development of CV disease (MI, stroke, heart failure, PAD)^a - Orthostatic hypotension is a risk factor for falls in older adults^b
Diastolic BP/5th Korotkoff sound	- Point at which pressure is low enough to return to laminar flow - Sound created by turbulent flow ceases	
Heart sounds	- Noises created in the heart by the turbulence in blood flow caused by the abrupt closing of the heart valves - The first heart sound (S1): closing of the aorto-ventricular (A-V) valve - The second sound (S2): closing of the pulmonary valve - A third, lower-pitched sound may occur - S4 occurs just after the end of diastole, before systole begins - Murmurs result from turbulent flow and usually are described as a 'whooshing' sound	- S1 and S2: normal heart sounds - S3: abnormal sound in adults; may indicate excess of blood volume in the heart, possibly signifying heart failure - S4: abnormal sound in adults; may indicate a failing left ventricle - Murmurs: often sign of mitral valve regurgitation or aortic valve stenosis
Bruits	- Sound created by turbulent flow in an artery - Caused by arterial narrowing - Blood flows faster at the narrowed point, causing turbulence and creating an audible sound	Bruits in the carotid or femoral artery or the aorta may signify significant stenosis requiring further diagnosis or intervention
Breath sounds	- Sounds generated as air moves through the respiratory system - In healthy individuals, breath sounds should be clearly audible throughout the lungs bilaterally	- Sounds that are not usually present (e.g., crackles, wheezes, friction rubs) may indicate a respiratory abnormality that requires treatment - Timing, pitch, and quality provide clinical clues about cause
Bowel sounds	- Assessment of intensity, frequency, and quality of bowel sounds should be divided into four quadrants - Sounds have different meanings depending on the quadrant from where they arise - Hypoactive bowel sounds are defined as one every three or four minutes	- Absence of bowel sounds can be confirmed only after listening for five minutes - A pattern of hyperactive bowel sounds in one location and hypoactive or absent bowel sounds in a lower location may indicate an obstruction - High-pitched or rushing noises may indicate an obstruction

Note: BP = blood pressure; CV = cardiovascular; MI = myocardial infarction; PAD = peripheral arterial disease.

^aGo et al. (2014).

^bShaw & Claydon (2014).

Table 7.3 Examples: research employing therapeutic use of sound or noise reduction in clinical care

Intervention	Example
Noise reduction	• Promotion of hearing protection devices in construction workers at risk of noise-induced hearing loss related to exposure to harmful noise^a • Noise reduction as part of a bundled intervention to improve sleep and reduce delirium in ICU patients^b • ICU improvement of alarm performance, reduction of unnecessary and false alarms^c • Review: effects of ambient stressors on sleep deprivation in ICU patients^d
Music therapy	• Patient-directed music intervention to reduce anxiety in mechanically ventilated ICU patients^e • Developmental and medical benefits of music therapy in preterm infants in the NICU^{f,g}

Note: ICU = intensive care unit; NICU = neonatal intensive care unit.

^aKerr, Lusk, & Ronis (2002).

^bPatel, Baldwin, Bunting, & Laha (2014).

^cGorges, Markewitz, & Westenskow (2009).

^dFontana & Pittiglio (2010).

^eChlan et al. (2013).

^fStandley (2012).

^gAllen (2013).

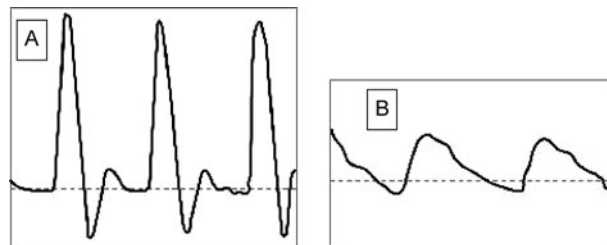


Figure 7.1 Ultrasound wave forms. Examples of normal (Panel A) and abnormal (Panel B) lower extremity arterial waveform patterns obtained using ultrasound. The normal Doppler arterial waveform is triphasic, with an initial steep upstroke, representing the high flow of systole. The second portion, a steep downstroke, indicates the reverse flow in early diastole. The third segment of the wave, a small peak, signifies the forward flow of late diastole. In contrast, Panel B illustrates the waveform seen at a point of arterial stenosis or occlusion changes to a biphasic or monophasic pattern. Uptake is slower, no reverse flow can be seen, and the second and third wave components are lost.

Webb, 2002, Chapter 3); it also increases protein synthesis (Tang, Guha, Tome, 2014). Ultrasound is also used noninvasively to destroy ureteral calculi, a procedure called lithotripsy (Xu et al., 2014).

Electricity and electromagnetism

Many physiological processes are driven by electrical and electromagnetic phenomena. All of matter is composed of subatomic particles, which have properties of electric charge. Negatively charged electrons are held to the nucleus of the atom by the positive charge of the

proton. An atom with equal numbers of protons and electrons is electrically neutral. Positive or negative ions are created by adding or removing electrons, producing electrical potential. The net movement of electrons is electrical current. Electrically charged particles emanate force and thus can act on other electrically charged particles from a distance through an electric force field. Electricity and magnetism are related in that a changing electric field creates a magnetic field and a changing magnetic field creates an electric field (Davidovits, 2013, Appendix B).

Measurement: electricity

Electric charge is measured in units called coulombs (C). The electric field is measured in units of volt per meter. Electrical current, the rate of electron flow, is measured in amperes. Resistance is the opposition to electrical current flow, measured in ohms (Davidovits, 2013, Appendix B).

Applications: electricity and electromagnetism

Common examples of electricity-based assessment technology include (a) electrocardiography (ECG), which records the electrical activity of the heart; (b) electroencephalography (EEG), which records electrical patterns of the brain; and (c) electromyography (EMG), which records the electrical activity of the skeletal muscles. Experts can diagnose disorders and dysfunction based on recognition of normal versus abnormal electrical wave form patterns in ECGs, EEGs, and EMGs (Davidovits, 2013, Chapter 14).

Magnetic resonance imaging (MRI) utilizes the proton spin that exists in molecules, which creates a small internal magnetic field (or axes/moments) (Davidovits, 2013, Chapter 17; Dillon, 2012, Chapter 1; Edelman & Warach, 1993). Images are produced by applying a strong external magnetic field, which causes the small internal magnetic fields with low energy states to align with the external magnetic field and the small internal magnetic fields atoms with high energy states to align against the magnetic field. The energy states are then equalized using a radiofrequency signal that matches the proton frequency, causing all the magnetic fields to align within the higher energy state (parallel). The signal is then “turned off” and some of the magnetic axes of the protons reequilibrate and align with the external magnetic field (parallel), releasing a signal (energy release) that can be detected. The time for this “relaxation” to take place is the T1 signal (Dillon, 2012; Edelman & Warach, 1993; Metter et al., 2000, Chapter 6; Webb, 2002, Chapter 4). The T2 relaxation signal is obtained during the time it takes for magnetic axes of the protons to reequilibrate and align against (become perpendicular to) the external magnetic field. Numeric algorithms are then used to analyze the signals and produce an image. Two- and three-dimensional images of tissue can be obtained by varying the radiofrequency signal, creating T1 or T2 weighted signals (Dillon, 2012, Chapter 1; Edelman & Warach, 1993; Metter et al., 2000, Chapter 6; Webb 2002, Chapter 4). The strength of the energy difference between atoms that can be detected (detected signal) depends upon the external magnetic field that is applied (the stronger the magnet, the larger the energy differences, and the stronger the potential signal), the proton density, and T1 and T2 weighted time differences. Functional MRI (fMRI) enables detection of metabolically active tissues. The blood oxygen level dependent (BOLD) effect created during fMRI allows neural activity to be inferred (Greve et al., this volume). For example, fMRI has been used to assess differences in brain responses in healthy weight to obese midlife women during a goal-directed decision task (Bosak & Martin, 2014).

Many therapeutic interventions use electrical impulses to achieve physiological benefits in a range of conditions, from dysrhythmias to paralysis. Examples are listed in Table 7.4.

Table 7.4 Examples: electrical devices used for therapeutic interventions

<i>Device</i>	<i>Method</i>	<i>Purpose</i>
Electronic pacemaker	Generates periodic electric impulses to the heart via electrodes	Initiates and controls heart rate
Cardiac defibrillator	Delivers approximately 13 amperes of current to tetanize the heart; when removed, normal rhythm can be restored	Restore normal rhythm during ventricular fibrillation, prevents sudden cardiac death
Hearing aid	Microphone converts sound to an electrical signal, which is amplified and converted back to sound	Enhance hearing
Electrical muscle stimulation	Small electrical impulse is delivered to muscles to simulate an action potential and cause muscles to contract	Prevent disuse atrophy in immobilized individuals after injury; increase tone of paralyzed muscles
Deep brain stimulation	Electrical stimulation to targeted areas in the brain that control movement, blocking abnormal nerve signals	Neurological symptoms – most commonly the debilitating motor symptoms of Parkinson's disease
Transcutaneous electrical nervous stimulation (TENS)	Biphasic pulsed currents in a repetitive manner using pulse durations of 50–250 ms and pulse frequencies of 1–200. Conventional TENS use low intensity, high frequency, but can also be high intensity, low frequency or high intensity, high frequency	Different TENS techniques are used to selectively activate groups of nerve fibers to elicit mechanisms leading to pain relief using the so called gate-control theory
Exoskeleton robotics ^a	Wearable robots that respond to the cognitive signal from the user of intent or volition to move	Allows humans to overcome physical and motor limits using artificial devices that help them use their limbs through artificial devices

^aPons (2010).

Force

Force is any external push or pull upon an object interacting with another object, causing it to change in movement, direction, or geometrical construction. Force can cause a flexible object to deform. The distribution of forces through the body is called mechanical stress. Mechanical stress is a physical quality that expresses the internal forces that neighboring particles of a continuous material exert on each other. Stress usually causes deformation of solid materials, or flow in fluids. Pressure is force divided by the area to which it is applied (Davidovits, 2013, Appendix A).

Several types of force can be exerted on solid tissues. Tensile force and compressive force work perpendicular to the stress plane, either pulling away from (tensile) or going toward (compressive) the center of the stress plane. Shear forces are unaligned forces that push or pull a body in opposite directions parallel to the stress plane (Figure 7.2).

Measurement: force and pressure

Force equals mass multiplied by acceleration ($F = M \times A$). The unit of measurement of force is the newton (NT); the dyne (dyn) is 10^5 NT, and is defined as the force it takes to accelerate a mass of one gram one centimeter per second per second. A unit of pressure (a pascal; Pa) is

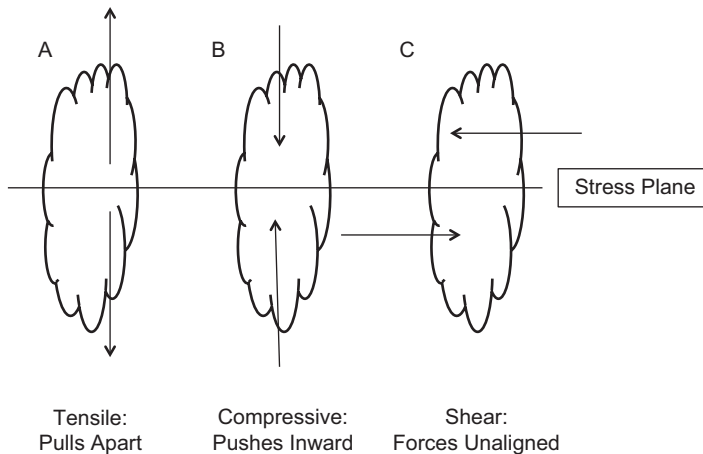


Figure 7.2 Forces on a solid object. Directions of force during tensile (Panel A), compressive (Panel B), and shear (Panel C) stress are shown.

defined as NT/m^2 . Pressure can also be measured by its ability to displace a column of liquid in a manometer, often measured in millimeters of mercury (mmHg). Blood pressure is most commonly reported in mmHg (“clinical units”); mercury sphygmomanometers are no longer used, but measurements from other methods are usually converted to mmHg (Babbs, 2012). A dynamometer measures force or power and is often used to measure muscle strength, including grip strength. Dynamometers have been shown to reliably measure muscle strength in a range of populations (Dekkers, Rameckers, Smeets, & Janssen-Potten, 2014; Williams, Kahn, & Randall, 2014; Vanpee, Hermans, Segers, & Gosselink, 2014).

Applications: force and pressure

Traction therapy is the use of tensile force to try to separate vertebrae or reduce skeletal fractures. Traction therapy has been beneficial in reducing displacement in patients with cervical spine trauma (Aebi, 2010), but evidence for efficacy for treatment of chronic low back pain is limited (Gay & Brault, 2008). Graded external compression therapy on the lower extremity uses compression force to oppose the elevated hydrostatic forces seen in venous hypertension; compression therapy in the form of graded compression stockings or wraps is recommended for all patients with chronic venous disease (Eberhardt & Rafetto, 2014). Shear stress contributes to skin breakdown. Shear stress arising from the friction of pulling skin against the bedding when repositioning a patient in bed creates sufficient shear force to cause skin breakdown particularly in older and less mobile individuals. Tissue compression, shear stress and wet skin combined provide the greatest risk for development of pressure ulcers (Sopher & Gefen, 2011). Prevention of shear stress in immobilized or frail individuals is essential to avoid skin breakdown.

Flow

Flow of material inside the human body is essential for biophysical functioning. Blood flowing through the circulatory system supplies tissues with oxygen and nutrients and allows for elimination of cellular waste. Cerebrospinal fluid protects and controls the environment of

the brain and spinal cord. Flow of lymphatic fluid is essential for elimination of cellular waste products. Flow of urine from the kidney provides a mechanism for elimination of waste and fluid. Large joints contain synovial fluid which lubricates surfaces and allows for range of motion. Interruption of flow has negative physiological consequences as with atherosclerosis, arthritis, lymphedema, hydrocephalus, and urinary retention. Understanding the biophysical properties of flow is a first step to understanding the physiological consequences of normal and abnormal flow of fluid in the human body. Arterial blood flow is used to illustrate key principles.

Blood flow is the *volume* of blood that passes a point per unit time, often expressed in milliliters per minute. Blood flow velocity is the *distance* that blood moves per unit of time, usually expressed in centimeters per second. Flow is equal to flow velocity multiplied by the area of a vessel. Hydrodynamic laws describe fluid flow. Hemodynamic flow is more complex than fluid flow under standardized experimental conditions, however, because fluid blood contains cells and other particles. Arterial flow is laminar (streamlined; flows in parallel layers); it is fastest at the center of the artery and significantly slower along the walls (Davidovits, 2013, Chapter 8; Guyton & Hall, 2006, Chapter 14), shown by Poiseuille's law:

$$Q = \pi r^4 (P_1 - P_0) / 8\eta L, \quad (7.5)$$

where Q is flow, r is the radius of the blood vessel, P_1 is the pressure at the intake of the vessel, P_0 is the pressure at the end of the vessel, η is the viscosity of the blood, and L is the length of the vessel segment (Figure 7.3). Poiseuille's law illustrates the importance of the radius of the blood vessel in the control of flow rate and pressure. Doubling the radius of the blood vessel results in a 16-fold increase in flow rate and, conversely, a very small reduction in the radius of the vessel markedly reduces flow rate (Guyton & Hall, 2006, Chapter 14).

Arterial flow is also affected by resistance to flow (vascular resistance), following Ohm's law:

$$Q = (P_1 - P_0) / R, \quad (7.6)$$

where Q is blood flow, P_1 is the pressure at the beginning of the vessel segment, P_0 is the pressure at the end of the vessel segment, and R is resistance. Resistance is primarily related to the radius of the vessel segment, and also to the segment length and blood viscosity (Dillon, 2012).

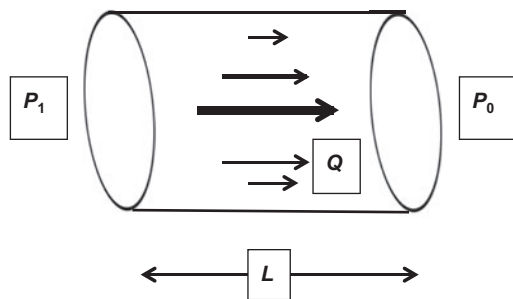


Figure 7.3 Laminar flow of blood with viscosity η through an arterial segment of length L with radius r . Note the parallel streamlining. As shown by the weight of the lines, flow is faster at the center of the artery than near the wall.

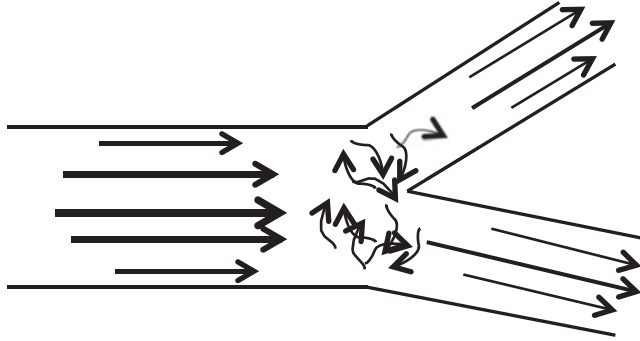


Figure 7.4 Turbulent flow of arterial blood at a point of bifurcation.

Poiseuille's and Ohm's laws assume that the radius of an artery is consistent throughout a segment under consideration. Whenever there is narrowing, the same volume of blood per second must flow through a smaller lumen. Thus, flow velocity increases. When velocity increases above a certain point, flow becomes turbulent (erratic in magnitude and direction) (see Figure 7.4). Turbulent flow occurs at the point of arterial narrowing caused by the atherosclerotic plaque and at bifurcations of the arterial tree. The Reynolds number (Re) is the ratio of inertial forces to viscous forces,

$$Re = \rho V D / \eta, \quad (7.7)$$

where ρ is fluid density, D is tube diameter, V is velocity, and η is viscosity; higher values of Re reflect the tendency of turbulence to occur (Dillon, 2012; Glaser, 2010). Turbulent flow can produce an atherogenic and prothrombotic environment and increases the likelihood of atherosclerotic plaque rupture when plaque is present (Chiu & Chien, 2011).

Measurement and applications: flow

Throughout the arterial tree, blood vessels branch into groups of smaller and smaller parallel vessels, from the aorta, through the arteries, arterioles, to the capillaries. The radius of an individual vessel becomes progressively smaller as blood flows through the arterial system, and yet within the individual vessel, rate of flow decreases, seemingly in contrast to Poiseuille's law. This is because the arterial resistance is distributed across all parallel vessels in a given arterial bed, increasing the total area, and reducing the rate of flow in individual vessels (Guyton & Hall, 2006, Chapter 14).

Arterial blood flow can be measured invasively and noninvasively. The electromagnetic flowmeter records the electromagnetic force in blood that is flowing through an electromagnetic field. Two poles of a strong magnet are placed on either side of the blood vessel; electrodes are placed perpendicular to the magnet. When blood flows through the vessel, an electrical voltage is generated and recorded using an electronic recording system. The electromagnetic flowmeter can measure very rapid changes in both steady and pulsatile flow (Guyton & Hall, 2006, Chapter 14).

Ultrasound is also used to assess arterial blood flow and velocity. By combining mean velocity measured via Doppler ultrasound and measuring the vessel diameter using B-mode ultrasound, the measurement of blood flow can be derived. This requires a trained operator and relatively expensive equipment. The use of ultrasound allows for detection of variations in velocity identifying areas of stenosis and other abnormal flow patterns (Cobbold, 2006).

Table 7.5 Examples: interventions to enhance flow of bodily fluids

<i>Approach</i>	<i>Method</i>	<i>Purpose</i>
Balloon angioplasty with or without stenting	• Invasive insertion of a catheter across an arterial stenosis, inflating a balloon to compress the plaque against the arterial wall to increase the lumen size • Can include deployment of a stent (a metal mesh tube) that adheres to the arterial wall to increase durability of the intervention	• To increase blood flow through an artery to ensure physiological viability of the tissue
Manual lymphatic drainage/complete decongestive therapy ^a	• Gentle massage to manually move the lymphatic fluid through the lymph vessels into the venous system for subsequent elimination	• Replaces the function of normal lymphatic flow that has been compromised due to damaged or nonfunctioning lymphatic vessels
External counterpulsation therapy (ECP) ^b	• Noninvasive therapy consisting of intermittent rapid cuff inflation and deflation to compress the arteries of the calves, thighs, and buttocks, synchronized to the patient's electrocardiogram • Arterial compression increases venous return to the heart, reduces peripheral vascular resistance, and reduces myocardial workload	• In individuals with angina or heart failure, ECP stimulates capillary opening and formation of collateral blood vessels, creating a 'bypass' around narrowed or blocked coronary arteries in patients who do not want or are contraindicated for traditional revascularization procedures
Venous compression pump therapy ^c	• Cuffs placed on the lower leg are programed to serially inflate and deflate from distal to proximal to mimic normal calf muscle pump functioning	• To increase venous return, minimize venous pooling, lower extremity edema, and venous ulceration

^aLasinski et al. (2012).

^bEstahbanaty et al. (2007).

^cFelty & Rooke (2005).

Intravascular ultrasound can be used to assess flow gradients such as coronary flow velocity reserve (ratio of peak flow velocity during peak hyperemia to baseline peak flow velocity) (Mohler, Gornik, Gerhard-Herman, Misra, Olin, & Zierler, 2012).

Flow can be altered by many interventions including exercise, positional changes, manual therapies, mechanical therapies, and invasive openings of narrowed arteries. Table 7.5 lists examples of interventions that are common to increasing flow.

Summary

This chapter has provided a brief overview of select biophysical principles, measurement of biophysical phenomena, and applications in nursing research. Cutting-edge research in nursing will require an understanding of biophysical principles and methods in order to fully investigate health-wellness processes and physiological phenomena.

References

- Adam Noah, J., Spierer, D. K., Gu, J., & Bronner, S. (2013). Comparison of steps and energy expenditure assessment in adults of Fitbit Tracker and Ultra to the Actical and indirect calorimetry. *Journal of Medical Engineering and Technology*, 37, 456–462.
- Aebi, M. (2010). Surgical treatment of upper, middle and lower cervical injuries and non-unions by anterior procedures. *European Spine Journal*, 19(Suppl. 1), S33–S39.
- Alkadri, M. E., Peters, M. N., Katz, M. J., & White, C. J. (2013). State-of-the-art paper: Therapeutic hypothermia in out of hospital cardiac arrest survivors. *Catheterization and Cardiovascular Interventions*, 82(4), E482–E490.
- Allen, K. A. (2013). Music therapy in the NICU: Is there evidence to support integration for procedural support? *Advances in Neonatal Care*, 13, 349–352.
- Alles, M. S., Eussen, S. R., & van der Beek, E. M. (2014). Nutritional challenges and opportunities during the weaning period and in young childhood. *Annals of Nutrition & Metabolism*, 64, 284–293.
- American Academy of Pediatrics, Committee on Environmental Health (1997). Noise: A hazard for the fetus and newborn. *Pediatrics*, 100, 724–727.
- American College of Sports Medicine. (2010). *Guidelines for graded exercise testing and exercise prescription* (8th ed.) Baltimore, MD: Lippincott, Williams & Wilkins.
- American Thoracic Society. (2002). ATS statement: Guidelines for the six-minute walk test. *American Journal of Respiratory Critical Care Medicine*, 166, 111–117.
- Babbs, C. F. (2012). Oscillometric measurement of systolic and diastolic blood pressures validated in a physiologic mathematical model. *BioMedical Engineering OnLine*, 11, 56. doi:10.1186/1475-925X-11-56
- Bisson, J., & Younker, J. (2006). Correcting arterial blood gases for temperature: (When) is it clinically significant? *Nursing in Critical Care*, 11, 232–238.
- Bosak, K., & Martin, L. (2014). Neuroimaging of goal-directed behavior in midlife women. *Nursing Research*, 63, 388–396.
- Bouchet, L., Sanford, L., Goodchild, G., Bova, F., Buatti, J., & Friedman, W. (2001). Calibration of three dimensional ultrasound images for image-guided radiation therapy. *Physics in Medicine and Biology*, 46, 559.
- Chen, T., & Chou, L. S. (2013). Altered center of mass control during sit-to-walk in elderly adults with and without history of falling. *Gait & Posture*, 38, 696–701.
- Chiu, J. J., & Chien, S. (2011). Effects of disturbed flow on vascular endothelium: Pathophysiological basis and clinical perspectives. *Physiological Reviews*, 91, 327–387.
- Chlan, L. L., Weinert, C. R., Heiderscheit, A., Tracy, M. F., Skaar, D. J., Guttormson, J. L., & Savik, K. (2013). Effects of patient-directed music intervention on anxiety and sedative exposure in critically ill patients receiving mechanical ventilatory support: A randomized clinical trial. *JAMA*, 309, 2335–2344.
- Chodera, J. D., & Mobley, D. L. (2013). Entropy-enthalpy compensation: Role and ramifications in biomolecular ligand recognition and design. *Annual Review of Biophysics*, 42, 121–142.
- Cobbold, R. S. C. (2006). *Foundations of biomedical ultrasound*. Oxford: Oxford University Press.
- Cotterill, R. (Ed.) (2008). *Biophysics*. West Sussex: Wiley.
- Davidovits, P. (2013). *Physics in biology and medicine* (4th ed.). Waltham, MA: Elsevier.
- de van der Schueren, M., Elia, M., Gramlich, L., Johnson, M. P., Lim, S. L., Philipson, T., . . . Prado, C. M. (2014). Clinical and economic outcomes of nutrition interventions across the continuum of care. *Annals of the New York Academy of Sciences*, 1321, 20–40.
- Dekkers, K., Rameckers, E., Smeets, R., & Janssen-Potten, Y. (2014). Upper extremity strength measurement of children with cerebral palsy: A systemic review of available instruments. *Physical Therapy*, 94, 609–622.
- Dillon, P. (Ed.) (2012). *Biophysics: A physiological approach*. Cambridge: Cambridge University Press.
- Dodd, S. R., Lancaster, G. A., Craig, J. V., Smyth, R. L., & Williamson, P. R. (2006). In a systematic review, infrared ear thermometry for fever diagnosis in children finds poor sensitivity. *Journal of Clinical Epidemiology*, 59, 354–357.
- Eberhardt, R. T., & Raftery, J. D. (2014). Chronic venous insufficiency. *Circulation*, 130, 333–346.
- Edelman, R. R., & Warach, S. (1993). Magnetic resonance imaging (1). *New England Journal of Medicine*, 328, 708–716.
- Estahbanaty, G., Samiei, N., Maleki, M., Noohi, F., Mohebi, A., Ojaghi, Z., . . . Soran, O. (2007). Echocardiographic characteristics including tissue Doppler imaging after enhanced external counterpulsation therapy. *American Heart Hospital Journal*, 5, 241–246.
- Felty, C. L., & Rooke, T. W. (2005). Compression therapy for chronic venous insufficiency. *Seminars in Vascular Surgery*, 18, 36–40.

- Fontana, C. J., & Pittiglio, L. I. (2010). Sleep deprivation among critical care patients. *Critical Care Nursing Quarterly*, 33(1), 75–81.
- Forbes, S. S., Eskicioglu, C., Nathens, A. B., Fenech, D. S., Laflamme, C., McLean, R. F., . . . Best Practice in General Surgery Committee, University of Toronto. (2009). Evidence-based guidelines for prevention of perioperative hypothermia. *Journal of the American College of Surgeons*, 209, 492–503.
- Freire, E. (2008). Do enthalpy and entropy distinguish first in class from best in class? *Drug Discovery Today*, 13, 869–874.
- French, S. D., Cameron, M., Walker, B. F., Reggars, J. W., & Esterman, A. J. (2006). A Cochrane review of superficial heat or cold for low back pain. *Spine*, 31, 998–1006.
- Gay, R. E., & Brault, J. S. (2008). Evidence-informed management of chronic low back pain with traction therapy. *Spine Journal*, 8, 234–242.
- Gillman, L. M., & Kirkpatrick, A. W. (2012). Portable bedside ultrasound: The visual stethoscope of the 21st century. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 20, 18. doi:10.1186/1757-7241-20-1
- Glaser, R. (Ed.) (2010). *Biophysics*. Heidelberg: Springer-Verlag.
- Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Blaha, M. J., . . . Turner, M. B. (2014). Heart disease and stroke statistics – 2014 update: A report from the American Heart Association. *Circulation*, 129, e28–e292.
- Gorges, M., Markewitz, B. A., & Westenskow, D. R. (2009). Improving alarm performance in the medical intensive care unit using delays and clinical context. *Anesthesia and Analgesia*, 108, 1546–1552.
- Gresham, E., Byles, J. E., Bisquera, A., & Hure, A. J. (2014). Effects of dietary interventions on neonatal and infant outcomes: A systematic review and meta-analysis. *American Journal of Clinical Nutrition*, 100, 1298–1321.
- Gurlanik, J. (2012). Assessing physical performance in the older patient. Retrieved from www.grc.nia.nih.gov/branches/leps/sppb/
- Guyton, A. C., & Hall, J. E. (2006). *Textbook of medical physiology* (11th ed.). Philadelphia, PA: Elsevier Saunders.
- Hammer, M. S., Swinburn, T. K., & Neitzel, R. L. (2014). Environmental noise pollution in the United States: Developing an effective public health response. *Environmental Health Perspectives*, 122, 115–119.
- Harris, T., Kerry, S., Victor, C., Ekelund, U., Woodcock, A., Iliffe, S., . . . Cook, D. (2013). Randomized controlled trial of a complex intervention by primary care nurses to increase walking in patients aged 60–74: Protocol of the PACE-Life (Pedometer Accelerometer Consultation Evaluation-Life) trial. *BMC Public Health*, 13(5). doi:10.1186/1471-2458-13-5
- Heermann, D. W. (2011). Physical nuclear organization: Loops and entropy. *Current Opinion in Cell Biology*, 23, 332–337.
- Holtzclaw, B. J. (1998). New trends in thermometry for the patient in the ICU [Review]. *Critical Care Nursing Quarterly*, 21(3), 12–25.
- Holtzclaw, B. J. (2001). Circadian rhythmicity and homeostatic stability in thermoregulation [Review]. *Biological Research for Nursing*, 2, 221–235.
- Hooper, V. D., & Andrews, J. O. (2006). Accuracy of noninvasive core temperature measurement in acutely ill adults: The state of the science. *Biological Research for Nursing*, 8, 24–34.
- Hughes, L. F. (2007). The fundamentals of sound and its measurement. *Journal of the American Association of Animal Laboratory Science*, 46, 14–19.
- Issels, R. D., Lindner, L. H., Verweij, J., Wust, P., Reichardt, P., Schem, B. C., . . . European Society for Hyperthermic Oncology (ESHO). (2010). Neo-adjuvant chemotherapy alone or with regional hyperthermia for localised high-risk soft-tissue sarcoma: A randomised phase 3 multicentre study. *Lancet. Oncology*, 11, 561–570.
- Kerr, M. J., Lusk, S. L., & Ronis, D. L. (2002). Explaining Mexican American workers’ hearing protection use with the health promotion model. *Nursing Research*, 51, 100–109.
- Keteyian, S. J., Fleg, J. L., Brawner, C. A., & Piña, I. L. (2010). Role and benefits of exercise in the management of patients with heart failure. *Heart Failure Reviews*, 15, 523–530.
- Lasinski, B. B., McKillip Thrift, K., Squire, D., Austin, M. K., Smith, K. M., Wanchai, A., . . . Armer, J. M. (2012). A systematic review of the evidence for complete decongestive therapy in the treatment of lymphedema from 2004 to 2011. *Physical Medicine and Rehabilitation*, 4, 580–601.
- Lesniak, B. P., Loveland, D., Jose, J., Selley, R., Jacobson, J. A., & Bedi, A. (2014). Use of ultrasonography as a diagnostic and therapeutic tool in sports medicine. *Arthroscopy*, 30, 260–270.
- Lillicrop, K. A., & Burdge, G. C. (2011). The effect of nutrition during early life on the epigenetic regulation of transcription and implications for human diseases. *Journal of Nutrigenetics and Nutrigenomics*, 4, 248–260.

- Madden, A. M., Parker, L. J., & Amirabdollahian, F. (2013). Accuracy and preference of measuring resting energy expenditure using a handheld calorimeter in healthy adults. *Journal of Human Nutrition and Dietetics*, 26, 587–595.
- Metter, R., Beutel, J., & Kundel, H. (2000). *Handbook of medical imaging*. Bellingham, WA: SPIE.
- Mifflin, M. D., St. Jeor, S. T., Hill, L. A., Scott, B. J., Daugherty, S. A., & Koh, Y. O. (1990). A new predictive equation for resting energy expenditure in healthy individuals. *American Journal of Clinical Nutrition*, 51, 241–247.
- Mohler, E. R., Gornik, H. L., Gerhard-Herman, M., Misra, S., Olin, J. W., & Zierler, R. J. (2012). Appropriate use criteria for peripheral vascular ultrasound and physiological testing. Part I: Arterial ultrasound and physiological testing. *Journal of the American College of Cardiology*, 60, 242–276.
- National Heart, Lung, and Blood Institute. (2013). Managing overweight and obesity in adults. Systematic evidence review from the Obesity Expert Panel, 2013. Retrieved from www.nhlbi.nih.gov/health-pro/guidelines/in-develop/obesity-evidence-review
- Novak, I. (2011). Effective home programme intervention for adults: A systematic review. *Clinical Rehabilitation*, 25, 1066–1085.
- Pacheco-Acosta, J. C., Gomez-Correa, A. C., Florez, I. D., Cortes, J. E., Velez, D., Gomez, J., . . . Arboleda, S. (2014). Incidence of nutrition deterioration in nonseriously ill hospitalized children younger than 5 years. *Nutrition in Clinical Practice*, 29, 692–697.
- Patel, J., Baldwin, J., Bunting, P., & Laha, S. (2014). The effect of a multicomponent multidisciplinary bundle of interventions on sleep and delirium in medical and surgical intensive care patients. *Anaesthesia*, 69, 540–549.
- Pinto, B. M., & Floyd, A. (2007). Methodologic issues in exercise intervention research in oncology. *Seminars in Oncology Nursing*, 23, 297–304.
- Pons, J. L. (2010, May/June). Rehabilitation exoskeletal robotics. *IEEE Engineering in Medicine and Biology Magazine*. doi:10.1109/MEMB.2010.936548
- Pusnik, I., & Drnovsek, J. (2005). Infrared ear thermometers – parameters influencing their reading and accuracy. *Physiological Measurement*, 26, 1075–1084.
- Ruppar, T. M., Conn, V. S., Chase, J. A., & Phillips, L. J. (2014). Lipid outcomes from supervised exercise interventions in healthy adults. *American Journal of Health Behavior*, 38, 823–830.
- Sallis, J. F., Haskell, W. L., Wood, P. D., Fortmann, S. P., Rogers, T., Blair, S. N., & Paffenbarger, R. (1985). Physical activity assessment methodology in the Five City Project. *American Journal of Epidemiology*, 121, 91–106.
- Seden, G. (2011). The thermal environment. In R. J. Martin, A. A. Fanaroff, & M. C. Walsh (Eds.), *Fanaroff & Martin's neonatal-perinatal medicine* (9th ed., pp. 555–569). St. Louis, MO: Elsevier Mosby.
- Sharma, S. (2011). *Biophysics in nursing*. New Delhi, India: Japee Brothers.
- Shaw, B. H., & Claydon, V. E. (2014). The relationship between orthostatic hypotension and falling in older adults. *Clinical Autonomic Research*, 24, 3–13.
- Singer, P., Hiesmayr, M., Biolo, G., Felbinger, T. W., Berger, M. M., Goeters, C., . . . Pichard, C. (2014). Pragmatic approach to nutrition in the ICU: Expert opinion regarding which calorie protein target. *Clinical Nutrition (Edinburgh, Scotland)*, 33, 246–251.
- Sloane, R., Snyder, D. C., Demark-Wahnefried, W., Lobach, D., & Kraus, W. E. (2009). Comparing the 7-day physical activity recall with a triaxial accelerometer for measuring time in exercise. *Medicine and Science in Sports and Exercise*, 41, 1334–1340.
- Smith, J., Alcock, G., & Usher, K. (2013). Temperature measurement in the preterm and term neonate: A review of the literature. *Neonatal Network*, 32(1), 16–25.
- Smith, J., & Woodcock, A. (2008). New developments in the objective measurement of cough. *Lung*, 186(Suppl. 1), S48–S54.
- Sopher, R. & Gefen, A. (2011). Effects of skin wrinkles, age and wetness on mechanical loads in the stratum corneum as related to skin lesions. *Medical and Biological Engineering & Computing*, 49, 97–105.
- Sprouse-Blum, A. S., Gabriel, A. K., Brown, J. P., & Yee, M. H. (2013). Randomized controlled trial: Targeted neck cooling in the treatment of the migraine patient. *Hawaii Journal of Medicine & Public Health*, 72, 237–241.
- Standley, J. (2012). Music therapy research in the NICU: An updated meta-analysis. *Neonatal Network*, 31, 311–316.
- Tang, J., Guha, C., & Tome, W. A. (2014). Biological effects induced by non-thermal ultrasound and implications for cancer therapy. *Technology in Cancer Research and Treatment*, 21. Advance online publication. doi:10.7785/tcr.2012.500407

- Thompson, P.D., Buchner, D., Piña, I.L., Balady, G.J., Williams, M.A., Marcus, B.H., . . . Council on Nutrition, Physical Activity, and Metabolism (Subcommittee on Physical Activity). (2003). Exercise and physical activity in the prevention and treatment of atherosclerotic cardiovascular disease. *Circulation*, 107, 3109–3116.
- Treat-Jacobson, D., Bronas, U.G., & Salisbury, D. (2013). Exercise. In M. Snyder, R. Lindquist, & M.F. Tracy (Eds.), *Complementary/alternative therapies in nursing* (7th ed., pp. 299–322). New York, NY: Springer.
- Vanpee, G., Hermans, G., Segers, J., & Gosselink, R. (2014). Assessment of limb muscle strength in critically ill patients: A systematic review. *Critical Care Medicine*, 42, 701–711.
- Webb, A.G. (2002). *Introduction to biomedical imaging*. New York, NY: Wiley-IEEE Press.
- Williams, G., Kahn, M., & Randall, A. (2014). Strength training for walking in neurologic rehabilitation is not task specific: A focused review. *American Journal of Physical Medicine and Rehabilitation*, 93, 511–522.
- World Health Organization. (2014). Obesity and overweight. Fact sheet no. 311. Retrieved from www.who.int/mediacentre/factsheets/fs311/en/
- Wycherley, T.P., Moran, L.J., Clifton, P.M., Noakes, M., & Brinkworth, G.D. (2012). Effects of energy-restricted high-protein, low-fat compared with standard-protein, low-fat diets: A meta-analysis of randomized controlled trials. *American Journal of Clinical Nutrition*, 96, 1281–1298.
- Xu, Y., Lu, Y., Li, J., Luo, S., Liu, Y., Jia, Z., . . . Jia, S. (2014). A meta-analysis of the efficacy of ureteroscopic lithotripsy and extracorporeal shock wave lithotripsy on ureteral calculi. *Acta Cirurgica Brasileira*, 29, 346–352.

PART III

Prediction and explanation

This page intentionally left blank

8

STRUCTURAL EQUATION MODELING

Sarah J. Schmiede and Angela D. Bryan

This chapter introduces basic tenets and uses of structural equation modeling (SEM). The term SEM denotes classes of models that include confirmatory factor analysis, structural equation models for observed variables, and latent variable structural equation models (Loehlin, 1992). Confirmatory factor analysis models specify measurement relationships among observed and latent variables; observed variable structural equation models specify theoretical relationships among variables without a measurement structure imposed, and full structural equation models specify relationships among latent variables. Use of SEM is widespread in nursing research because it is suited to measurement problems (e.g., Folse, 2007; Lynn, Morgan, & Moore, 2009; Sousa, Ryu, Kwok, Cook, & West, 2007) and theory building (e.g., Melnyk, Crean, Feinstein, & Fairbanks, 2008; Phillips & Stuijbergen, 2009; Spence Laschinger & Leiter, 2006) and software is readily available. We present SEM models suitable for continuously varying data, discuss estimation and model assessment, introduce several special topics, provide an overview of common software packages, and list additional resources.

Terminology

In SEM, measured variables are distinguished from latent variables. *Measured variables* are manifest and directly observed. *Latent variables*, or constructs, are inherently unobservable, and their existence is inferred on the basis of relationships among measured variables, called indicators. Relationships may be correlational or directional. *Exogenous* observed or latent variables are analogous to predictors in multiple regression and *endogenous* observed or latent variables are analogous to regression outcomes; however, endogenous variables in SEM can serve both as outcomes and predictors of other variables depending upon their placement in a complex model. *Parameters* are constants governing or summarizing relationships among the variables in a theoretical system; they are usually unknown and the goal of an analysis is to obtain parameter estimates with known statistical properties based on relevant data. In some cases, parameters may be fixed at a specified value; it is also possible to impose equality or order constraints when justified by theory.

Schematics (path diagrams) are used to represent measured and latent variables and relationships among them in SEMs (McDonald & Ho, 2002). Following convention, measured variables are represented by rectangles; latent variables are represented by circles; and relationships are represented by curved double-headed or straight single-headed arrows (for correlational and

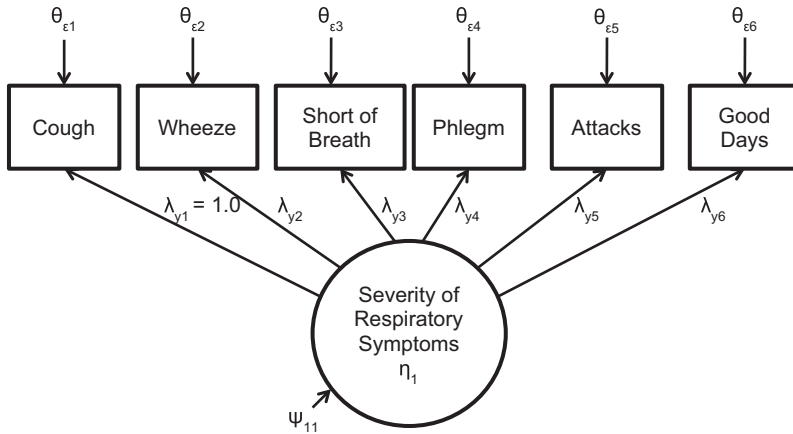


Figure 8.1 Example of a one-factor confirmatory factor analysis model. The factor loadings λ_{y1} to λ_{y6} are the elements of the Λ matrix and represent the regression of the indicators on the factor (η_1). The indicator uniquenesses are represented by θ_{ϵ_1} to θ_{ϵ_6} and comprise the diagonal elements of Θ_{ϵ} . The variance of the factor is represented by Ψ_{11} , the only element of the Ψ matrix for the single-factor model. The scale of the factor is set by constraining the first factor loading, λ_{y1} , to 1.0 as the *marker variable*. Note that any one of the other indicators could have been chosen as the marker variable, or Ψ_{11} could have been constrained to 1.0 to set the scale for the factor.

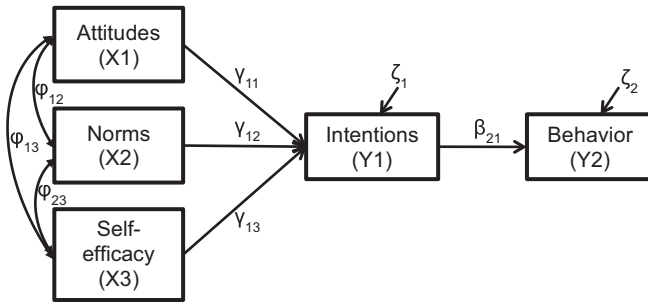


Figure 8.2 Example of an observed variable path model based on the theory of planned behavior (Ajzen & Madden, 1986). Covariances among the exogenous variables (attitudes, norms, self-efficacy) shown by the double-headed curved arrows and collected in the Φ matrix are fixed at sample values. Path coefficients linking the three exogenous variables directly to the endogenous intentions variable are collected in the Γ matrix and the direct link between the two endogenous variables is the appropriate element of $\mathbf{B}$. Unexplained variance in the endogenous variables is indicated by ζ_1 and ζ_2 .

directional relationships, respectively). Figures 8.1, 8.2, and 8.3 show schematics for a confirmatory factor analysis model, a path model for observed variables, and a latent variable path model. Each schematic is a representation of the respective analytic model. Analytic representations use matrix notation (Bollen, 1989, Chapter 2; Fox, 2009). Table 8.1 summarizes useful notation for SEM based on LISREL syntax (Jöreskog & Sörbom, 2006) that we follow here.

Statistical theory for SEM is based on properties of the *variance-covariance matrix*, a square, symmetric matrix where diagonal elements are variances of each of p measured variable and the $1/2p(p-1)$ non-duplicated off-diagonal elements are covariances among measured

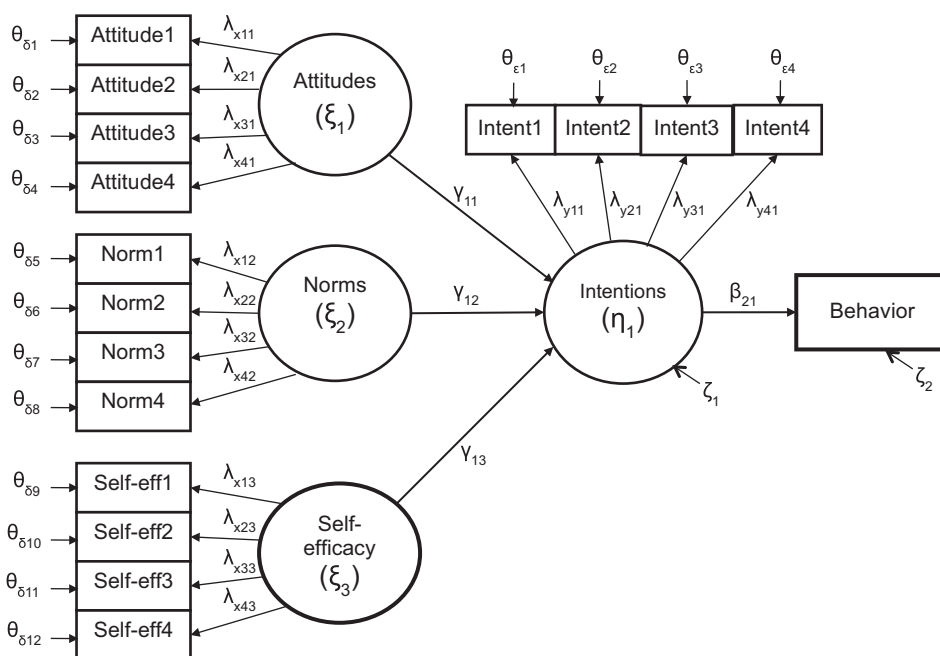


Figure 8.3 Example of a structural equation model with latent variables. Attitudes, norms, and self-efficacy are represented as exogenous latent factors (ξ), and intentions as an endogenous latent factor (η). Each factor has multiple indicators. Structural relationships link latent variables rather than observed variables (compare Figures 8.2 and 8.3). The inclusion of behavior in the model (represented as a rectangle) demonstrates the possibility of incorporating observed variables and latent factors in the same model.

Table 8.1 SEM matrix notation (following LISREL)

Matrix	Typical Element ^a	Name	Interpretation in a Full SEM ^b
Λ_x	λ_x	Lambda-x	Factor loadings for exogenous variables
Θ_δ	θ_δ	Theta-Delta	Uniquenesses for observed exogenous variables
Φ	ϕ	Phi	Variances and covariances for latent exogenous variables (factors)
*	ξ	Ksi	General name for latent exogenous variables
Λ_y	λ_y	Lambda-y	Factor loadings for endogenous variables
Θ_ϵ	θ_ϵ	Theta-Epsilon	Uniquenesses for endogenous variables
Ψ	ψ	Psi	Variances and covariances for latent endogenous variables (factors)
*	η	Eta	General name for latent endogenous factor
*	ζ	Zeta	Disturbance term for a latent endogenous variable
Γ	γ	Gamma	Path coefficient from exogenous to endogenous variables
B	β	Beta	Path coefficient from one endogenous variable to another

Notes

^aTypical elements are then subscripted with their location (row, column) in matrix.

^bInterpretation varies when other models are specified.

*These entries represent general names for the latent factors or the disturbance of an endogenous factor; they are not matrices.

variables. The $p \times p$ population matrix is designated Σ and the corresponding sample matrix is $\mathbf{S}$. The correlation matrix is the standardized form of the variance-covariance matrix and can be used to estimate some types of models in the SEM class (Cudeck, 1989; Jöreskog, 1978).

Confirmatory factor analysis model

In self report and other instruments, multiple items are used to represent a certain construct (e.g., quality of life, symptoms, self-efficacy), and relationships among responses to the items summarized in the covariances are used to infer underlying constructs (latent variables, factors). These theory-driven measurement relationships are specified using a confirmatory factor analysis model (CFA). The number of factors m , relationships between each of the p indicators and each of the m factors, relationships among factors, and uniquenesses are all part of the CFA model shown in Figure 8.1. The analytic model for scores $\mathbf{y}$ is

$$\mathbf{y} = \Lambda_y \boldsymbol{\eta} + \boldsymbol{\epsilon}, \quad (8.1)$$

where $\mathbf{y}$ contains the responses to items; Λ_y contains the regression coefficients of observed variables on factors, some of which are set to 0 when more than one factor is specified in the model; $\boldsymbol{\eta}$ are the latent variables; and $\boldsymbol{\epsilon}$ are item unique errors of measurement. This means that observed variables are composite variables. The latent variables specified in $\boldsymbol{\eta}$ are predictors, and the $\boldsymbol{\epsilon}$ term represents the part of $\mathbf{y}$ not predicted by the latent variables. The $\boldsymbol{\epsilon}$ term is composed of specific and random portions of the observed $\mathbf{y}$ variables. It is not possible to estimate parameters in Λ_y and $\boldsymbol{\epsilon}$ directly because scores on the latent variables are obviously not known. Nevertheless, estimates of the coefficients in Λ_y and the variances of $\boldsymbol{\epsilon}$ can be obtained using the model for the covariances among the observed p measured variables, such that

$$\Sigma = \text{cov}(\mathbf{y}) = \Lambda \Psi \Lambda^T + \text{diag}(\Theta_\epsilon), \quad (8.2)$$

where Σ is the population variance-covariance matrix, $\text{cov}(\mathbf{y})$ indicates the variances and covariances among measured variables, Λ is the $(p \times m)$ matrix of factor loadings, Ψ is the $(m \times m)$ matrix of correlations (or covariances) among factors, and $\text{diag}(\Theta_\epsilon)$ is the $(p \times p)$ diagonal matrix of unique variances for the measured variables. The Ψ matrix is set to a diagonal matrix when factors are hypothesized to be uncorrelated. Figure 8.1 illustrates a one-factor model for severity of respiratory dysfunction, represented by six symptom items. Parameters are shown on the schematic.

The logic behind CFA is that constructs are of prime scientific interest and individual indicators of them are never error-free representations of the factors. Thus, part of every indicator reflects the construct of interest; factors explain covariation among the indicators; and unique aspects of each indicator are a combination of specific and random variation (Bollen, 1989, p. 233). In CFA, the investigator hypothesizes the number of factors m and the pattern of factor loadings in Λ a priori and then places at least m^2 restrictions on factor loadings, variances, covariances, and uniquenesses. The set of scientifically interpretable parameters and the form of the model implies an estimated population covariance matrix Σ (Bollen, 1989, pp. 236–238) that is tested for fit against the observed data summarized in $\mathbf{S}$. Competing theories about the nature of related constructs and findings from past studies may suggest various measurement models, in which case selection of the best model from a set of scientifically plausible alternatives is a critical part of the assessment (Jöreskog, 1993).

The CFA model is flexible and can incorporate forms reflecting a wide range of hypotheses about measurement relationships. Models for sets of parallel and tau-equivalent test scores reflect more restrictive relationships among indicators than does the commonly estimated model for congeneric measurement (Jöreskog, 1971; see Sijtsma, this volume for a discussion of parallel tests). Higher-order factors can be specified to account for correlations (or covariances) in the Ψ matrix (Rindskopf & Rose, 1988). CFA models for multitrait-multimethod matrices (Kenny & Kashy, 1992) can sometimes be estimated (Grayson & Marsh, 1994). Bi-factor models specify a general factor and specific factors and are useful for exploring unidimensionality of an item set prior to estimating item response theory-based models (Baldwin & Wainer, this volume; Gibbons & Hedecker, 1992) and serve as a useful comparison to other proposed structures (Chen, West, & Sousa, 2006). Simultaneous estimation of CFA models in multiple groups is used to explore measurement invariance across populations (e.g., Sousa, West, Moser, Harris, & Cook, 2012).

Observed variable path models

Structural equation models using observed variables are extensions of multiple regression whereby a model reflecting a network of theory-based relationships among exogenous and endogenous measured variables is proposed; this class of models arose from early work in modeling genetic and environmental effects on phenotypes (Wright, 1921). An example of a path model is shown in Figure 8.2. The analytic model is given by

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \mathbf{\Gamma}\mathbf{x} + \boldsymbol{\zeta}, \quad (8.3)$$

where $\mathbf{y}$ is a vector of measured outcomes, $\mathbf{B}$ is a matrix of coefficients for effects of outcome variables on other outcomes, $\mathbf{x}$ is a vector of exogenous variables, $\mathbf{\Gamma}$ is a matrix of coefficients for the regression of $\mathbf{y}$ on $\mathbf{x}$, and $\boldsymbol{\zeta}$ contains the disturbances (error in prediction) for each of the outcomes.

Under the model, the implied population structure for $\boldsymbol{\Sigma}$ is a supermatrix (matrix whose elements are other matrices) with diagonal and lower triangular elements defined by

$$\boldsymbol{\Sigma}_{yy} = (\mathbf{I} - \mathbf{B})^{-1}(\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^T + \mathbf{\Psi})(\mathbf{I} - \mathbf{B})^{-1^T}, \quad (8.4a)$$

$$\boldsymbol{\Sigma}_{yx} = \mathbf{\Phi}\mathbf{\Gamma}^T(\mathbf{I} - \mathbf{B})^{-1^T}, \text{ and} \quad (8.4b)$$

$$\boldsymbol{\Sigma}_{xx} = \mathbf{\Phi}, \quad (8.4c)$$

where $\mathbf{\Phi}$ are covariances among the x variables and $\mathbf{\Psi}$ contains unexplained variances for the y variables (Bollen, 1989, p. 104, Equation 4.56). An observed variable path model based on the theory of reasoned action (Ajzen & Madden, 1986) shown in Figure 8.2 is typical; parameters are shown on Figure 8.2.

Path models may be recursive or non-recursive. *Recursive* models do **not** have feedback loops (such as $A \rightarrow B, B \rightarrow C$, and $C \rightarrow A$), reciprocal causes (such as $A \rightarrow B$ and $B \rightarrow A$), or correlations among disturbance terms. *Non-recursive* models may include loops, reciprocal effects, or correlated residual disturbances; strong theoretical rationale and challenging technical conditions must be met to estimate and interpret non-recursive models; see Kaplan, Harik, and Hotchkiss (2001), for example.

Path models are evaluated in terms of scientific interpretability, overall fit of the model, and the significance, direction, and magnitude of the path coefficients. Standardized and unstandardized

coefficients can be obtained and interpreted. Assumptions are that observations are independent, error (unexplained variance) is uncorrelated with predictors, and constructs are measured reliably. Furthermore, the model is assumed to be correctly specified with no key variables omitted and where the hypothesized paths between variables are meaningful, theoretically driven, and in the correct direction. The estimator must be correctly specified for the data for statistical properties to hold (Browne, 1984).

Structural equation models with latent variables

The measurement model specifies how indicators relate to latent constructs and allows correlations, but not directional relationships, to exist among factors. Path analysis refers to testing directional relationships among observed variables. Full structural equation models with latent variables combine these two submodels and incorporate directional relationships among latent factors as well.

As an example, Figure 8.3 shows a full SEM with latent variables arising from the theory of planned behavior. Again following LISREL notation, analytic specification involves a set of models including measurement models for both exogenous (x) variables and endogenous (y) variables and a theoretically justified path model for predicting endogenous factors (η) from other endogenous factors and from exogenous factors ξ such that

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon}, \quad (8.5a)$$

$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}, \text{ and} \quad (8.5b)$$

$$\boldsymbol{\eta} = \mathbf{B} \boldsymbol{\eta} + \boldsymbol{\Gamma} \boldsymbol{\xi} + \boldsymbol{\zeta}, \quad (8.5c)$$

where $\mathbf{\Lambda}_y$ contains coefficients for the regression of $\mathbf{y}$ on $\boldsymbol{\eta}$ (factor loadings for y variables), $\boldsymbol{\epsilon}$ contains unique components in $\mathbf{y}$, $\mathbf{\Lambda}_x$ contains coefficients for the regression of $\mathbf{x}$ on $\boldsymbol{\xi}$ (factor loadings for x variables), $\boldsymbol{\delta}$ contains unique components in $\mathbf{x}$, $\mathbf{B}$ is the matrix of coefficients for the regression of $\boldsymbol{\eta}$ on $\boldsymbol{\eta}$, $\boldsymbol{\Gamma}$ is the matrix of coefficients for the regression of $\boldsymbol{\eta}$ on $\boldsymbol{\xi}$, and $\boldsymbol{\zeta}$ contains the structural disturbance terms (residuals). Although the full SEM is complex, both the analytic representation in Equation 8.5 and the schematic (Figure 8.3) break it down into logical components that reflect theoretical measurement and structural relationships.

When the full SEM model holds and additional assumptions detailed by Jöreskog and Sörbom (2006; also see Browne, 1984) are met, the implied population structure for the covariance matrix among the observed y and x variables is again a supermatrix and has the form

$$\boldsymbol{\Sigma}_{yy} = \mathbf{\Lambda}_y (\mathbf{I} - \mathbf{B})^{-1} (\boldsymbol{\Gamma} \boldsymbol{\Phi} \boldsymbol{\Gamma}^T + \boldsymbol{\Psi}) [(\mathbf{I} - \mathbf{B})^{-1}]^T \mathbf{\Lambda}_y^2 \boldsymbol{\Theta}_\epsilon, \quad (8.6a)$$

$$\boldsymbol{\Sigma}_{yx} = \mathbf{\Lambda}_y (\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\Gamma} \boldsymbol{\Phi} \mathbf{\Lambda}_x^T, \text{ and} \quad (8.6b)$$

$$\boldsymbol{\Sigma}_{xx} = \mathbf{\Lambda}_x \boldsymbol{\Phi} \mathbf{\Lambda}_x^T + \boldsymbol{\Theta}_\delta. \quad (8.6c)$$

Full SEMs are evaluated in ways similar to both CFA and path models. Overall model fit, and significance and magnitude of individual parameters are important considerations. In an influential paper, Anderson and Gerbing (1988) argued for sequential estimation of first the measurement model and then the full structural equation model.

Causal inference and equivalent models

Use of the term “causal model” in the historical SEM literature has created mistaken notions about inferring cause from path models and structural equation models. The term “causal model” is generally a misnomer for this class of analytic techniques (e.g., Bollen, 1989; Hoyle, 1995; Loehlin, 1992), though there are dissenting voices (Pearl, 2012).

Equivalent models are models that have different parameters but imply the same variance-covariance matrix and exhibit identical fit to the same sample data (MacCallum, Wegener, Uchino, & Fabrigar, 1993). Equivalent models may exist for published models and may be particularly likely when design is cross-sectional or non-experimental. Researchers should be cognizant that equivalent models exist and should critically evaluate whether any of them are theoretically justified (MacCallum et al., 1993). Clear theoretical arguments for the chosen model should be provided.

Model identification and estimation

Parameters of SEM models are estimated by solving systems of equations. (See Fox, 2009, pp. 31–37 for a discussion about solving linear systems of equations.) SEM models must be “identified” in order to be estimated (Bollen, 1989). A model is identified if a unique solution for estimates of all the unknown model parameters exists. A necessary, but not sufficient, condition is that the number of knowns exceeds the unknowns. The knowns are based on the $p(p + 1)/2$ non-duplicated elements of the sample variance-covariance matrix. The unknowns are the q model parameters (e.g., factor loadings, path coefficients, and variance terms). The degrees of freedom (df) is then calculated as the difference, $p(p + 1)/2 - q$. If the number of knowns equals the number of unknowns, the df is zero and the model is “just identified.” If the knowns are less than the unknowns, the model is “underidentified” and cannot be estimated. If there are more knowns than unknowns, the model is “overidentified,” the df will be greater than 0, and a unique solution may be determined if other conditions are also met (Bollen, 1989; Skrondal & Rabe-Hesketh, 2004). Error messages are sometimes the first indication of non-identification, particularly in cases of empirical underidentification (Kenny & Milan, 2012) where the model meets the rules for identification at face value but there are issues in the sample data that prevent minimization of the fit function to a unique solution for every parameter. Overidentified models are desirable because model fit can be evaluated and meaningful conclusions can be drawn about whether the data are well represented by the hypothesized model.

Estimation is the process of determining values of population parameters based on sample data, using a method that produces estimates with known sampling properties (Stuart & Ord, 1991). In SEM, parameter estimates are obtained by minimizing a discrepancy between the sample variance-covariance matrix $\mathbf{S}$ and the variance-covariance matrix $\Sigma(\hat{\gamma})$ implied by the model (where $\hat{\gamma}$ is a vector containing the estimates for all the parameters in the model). It is not possible to obtain values of the estimates with direct hand calculation. Instead, numerical methods are used and estimation is iterative. For a given model with reasonable start values, a series of parameter estimates is obtained, such that a discrepancy between the sample matrix and the implied matrix decreases at each step until a convergence criterion is met (Cudeck, Klebe, & Henly, 1993; Lei & Wu, 2012).

Estimators for SEM are asymptotically justified. This means that theoretical properties of the estimators are known only for “large” samples, and that the small sample behavior of the estimates is not known. Optimally, the estimator will be asymptotically consistent, unbiased, and efficient (Lei & Wu, 2012). *Asymptotically consistent* means that as sample sizes approach infinity, the estimator should produce estimates that approach the true population parameter. An *unbiased* estimator is one whose

expected value for a particular parameter (averaged over an infinite number of independent random samples from a population) equals the value of that parameter in the population. The concept of *efficiency* refers to the sampling variability of the estimator; those with the smallest standard error are preferred. Note that there may be multiple consistent estimators that differ in efficiency.

Many estimators are available. Maximum likelihood (ML) is the most widely used estimator and the default of most programs. A closely related estimator is generalized least squares (GLS), which is similar to ML in assuming multivariate normality but may have advantages in small samples (Bentler, 2006). Both ML and GLS are popular in SEM applications because under conditions of multivariate normality of the input variables and a correctly specified model, they are consistent, unbiased, and efficient. When data are non-normally distributed, however, estimates of model fit based on the χ^2 test statistic and estimated parameter precision (standard errors) may be distorted (Curran, West, & Finch, 1996; Hoogland & Boomsma, 1998). The asymptotically distribution free (ADF) estimator can be used in situations of multivariate non-normality (Browne, 1984), although simulation work demonstrates large samples may be required. When distributions are heavy-tailed but symmetric, corrections to the standard error and χ^2 values are available (Satorra & Bentler, 1994). These corrections are found in several SEM programs and provide reasonable estimates over the usual range of sample sizes found in nursing (e.g., around 200 or more; Curran, West, & Finch, 1996). A variant of ML known as full-information maximum likelihood (FIML) is available for situations where data are either missing at random or missing completely at random (Arbuckle, 1996). FIML modifies the estimation procedure to make use of all available data, and is considered state-of-the-art for the treatment of missing data (Schafer & Graham, 2002). Newer estimators have been developed for data that are both non-normal and missing (Yuan & Bentler, 2000).

Choosing the appropriate estimator is crucial for proper assessment of fit of the model to the observed data and assuring that theoretical properties of the estimates and their standard errors obtain. The key challenge is to ensure that the estimator is correctly specified for the distribution of the data (Browne, 1984). This is not trivial, because estimators are not in general robust to violation of assumptions, normality may be uncommon (e.g., Micceri, 1989), and it is difficult to verify the distribution of the observations. Guidance about the finite sample behavior of the estimators can be obtained from results of sampling experiments (e.g., Hoogland & Boomsma, 1998). Multivariate kurtosis of the observations should be evaluated prior to using the normal theory ML and GLS estimators (Browne, 1984, p. 63).

Model assessment

Once model parameters have been estimated, the next step is to evaluate “fit” of the model to observed data. The parameter estimates for the model are used to compute the implied variance-covariance matrix among the observed variables *if the model was true* (West, Taylor, & Wu, 2012). The assessment of model fit answers the question of how closely the model-implied variance-covariance matrix $\Sigma(\hat{\gamma})$ matches the variance-covariance matrix $\mathbf{S}$ from the observed data. We address perspectives on model evaluation, evaluation of overall model fit, and statistical and scientific assessment of parameter estimates.

Perspectives on overall model assessment

The issue of model assessment is a source of considerable theoretical debate in the SEM literature, and different perspectives have arisen regarding the most appropriate and defensible strategy (Antonakis, Bendahan, Jacquart, & Lalive, 2013; Bentler & Bonnet, 1980; Hayduk & Glaser, 2000;

Marsh, Hau, & Wen, 2004; MacCallum, Roznowski, & Necowitz, 1992). Further complicating the issue, there are at least 40 known “fit indices” currently available (Marsh, Hau, & Grayson, 2005), so knowing which to rely on under which circumstances can be challenging.

Jöreskog (1993) offered a distinction among three research situations under which a model might be evaluated, and suggested that there might be differences in the most appropriate strategy for model assessment depending upon the goal of model testing. These situations are *strictly confirmatory*, *alternative models*, and *model generating*. In this section, we discuss the use of the indices most commonly seen in the literature, organizing our discussion using Jöreskog’s framework. See West, Taylor, and Wu (2012) for another review.

The strictly confirmatory case: Model testing

In the strictly confirmatory case, the researcher has formulated a model on the basis of a theory and is interested only in testing that particular model. Specifically, the researcher seeks to test the null hypothesis that the model is an adequate representation of the relationships in the population. The χ^2 test statistic is used to formally test the null hypothesis that $\Sigma = \Sigma(\gamma)$; this null hypothesis states that the covariance matrix in the population is equal to the matrix implied by the model, or equivalently, that the model is a “perfect fit” in the population. The χ^2 test statistic is equal to $(N - 1)$ times the minimum value of the fit function, with $df = 1/2p(p + 1) - q$. If the observed χ^2 exceeds the critical value for a given df and Type 1 error rate (often $\alpha < .05$), the interpretation is that the null hypothesis must be rejected. Unlike familiar hypothesis testing situations (e.g., tests about equality of means in two populations), *non-significant χ^2 values thus indicate reasonable fit (support for the scientific hypothesis) whereas significant χ^2 values represent lack of fit (rejection of the scientific hypothesis)*.

The validity of the χ^2 test rests on a number of assumptions, including that data from which the model was generated is a random sample of the population, a highly unusual circumstance in many areas of application. Further, like all hypothesis testing protocols, significance of the test is affected by sample size and the χ^2 is likely to be significant with a reasonable sample size. Nevertheless, if the purpose of the research is to accept or reject a single model, the χ^2 test is the only fit index that provides a test of the null hypothesis that allows for such a statistical decision (cf., Antonakis et al., 2013). An example when such a test would be of interest would be the evaluation of measurement equivalence across groups (cf., Reise, Widaman & Pugh, 1993); in this case, the researcher is truly interested in the test of the null hypothesis that measurement properties are the same (or not) in the groups.

The alternative models case: Selection of the best model

In model selection, a researcher has several competing models they wish to test, and seeks to compare those models in a single dataset. There are two possibilities regarding the relationship between these competing models. The models may be “nested,” meaning that parameters estimated in one model are a subset of the parameters estimated in the other model. In this case, models can be directly compared on the basis of subtracting the likelihood ratio χ^2 value obtained from the more restricted model from the likelihood ratio χ^2 value from the less restricted model, resulting in a likelihood ratio χ^2 -difference test ($\Delta\chi^2$) with degrees of freedom equal to the difference in degrees of freedom between the two models being tested. A non-significant $\Delta\chi^2$ test indicates that the models are an equivalent fit to the data. When models are not nested, they can be compared with fit indices that include criteria for the assessment of parsimony (e.g., Akaike’s Information Criterion [AIC]; Akaike, 1987).

The model generating case: Data-based model development

Perhaps the most common use of SEM is the situation in which a researcher has an initial idea, usually generated on the basis of theory and substantive knowledge, that results in a hypothesized model structure. In this case, rather than a simple up or down decision, the researcher seeks to discover the model that captures the best representation of the relationships among the constructs. Thus the null hypothesis of perfect fit that results from the likelihood ratio χ^2 test is a rather unrealistic bar to set. To quote Box (1979, p. 202), “all models are wrong, some are useful,” and the goal of model generation should be to evaluate how useful a particular model may be. The problems with the likelihood ratio χ^2 test empirically and conceptually for the model generation enterprise were almost immediately recognized and led to efforts to develop alternative assessments of model fit.

Absolute and comparative fit

Hu and Bentler (1999) noted that some analysts categorize fit indices along the broad dimension of whether they assess absolute or comparative fit. Absolute fit indices are functions of either the χ^2 test statistic or the residuals (difference between $\Sigma(\hat{\gamma})$ and **S**). Comparative fit indices evaluate the improvement in fit of an estimated model in comparison to some baseline model (Bentler, 1990). The default baseline model implemented in most software packages is the independence model, which estimates a variance for each measured variable but sets covariances to 0 (i.e., each observed variable is assumed to be independent of every other observed variable).

Goodness- and badness-of-fit

The various scalings of the fit indices results in some indicating goodness-of-fit and others badness-of-fit. All comparative fit indices are goodness-of-fit indices, which means that as the value of the index increases, the fit of a more highly parameterized nested model is “better” than the comparison baseline model. In contrast, absolute fit indices can be either goodness-of-fit indices or badness-of-fit indices (as the index increases, fit is worse).

Coming to a conclusion about model fit

Determining how information in the indices can be used to support an argument for or against some given model is not straightforward. While researchers have utilized “cut-offs” for the determination of a good-fitting model for each of these indices, the SEM literature in general recommends a flexible approach to evaluating overall model fit that includes the use of multiple indices, rather than a strict reliance on one index or the rejection of a model based on arbitrary cut-off values that may not be relevant to the research context (West et al., 2012; Marsh, Hau, & Wen, 2004). Further, it is important to also examine standardized residuals in addition to overall measures of fit, as residuals give clues to the sources of lack of fit in the case that fit is suboptimal in ways that measures of overall fit do not (Jöreskog, 1993). Ultimately, Box’s observation that models are optimally a useful representation of relationships among constructs is at the core of the model development idea. Our recommendation in deciding which model is the preferred representation is that the researcher not simply conclude that the chosen model meets some arbitrary cut-off on a particular fit index, but rather “to state clearly the criteria that are used in the comparison, in conjunction with descriptions of the models, characteristics of the data, and the purpose for which the models were constructed” (Cudeck & Henly, 1991, p. 518). It is within this important context that the selected model can best be understood.

Assessment of parameter estimates

Each estimated model parameter is represented by a point estimate and a standard error obtained as part of the overall estimation procedure. From the perspective of theory testing and theory development using SEM, the size, direction, significance, and scientific meaningfulness of parameter estimates is as important as overall fit. It is possible, though not common, that a model provides adequate fit to the data, while none of the hypothesized structural paths in the model is significant. This would be a clear sign that the model was not supportive of the researcher's hypotheses regarding the relationships among model constructs, despite an overall assessment that the model was a reasonable representation of the variance-covariance matrix. More commonly, some of the parameters are significant and some are not, and the researcher is left to make decisions about the relative degree to which hypotheses are supported. When fit is not consistent and/or parameters are not as expected, the researcher must make decisions regarding the inclusion or exclusion of constraints to the model. Such modifications should be undertaken with great caution, as deviation from the hypothesized model immediately moves the researcher from confirmatory to exploratory analysis (West et al., 2012) and would not, for example, be appropriate in a strict model testing situation. In contrast, the search for ways to change the model in order to improve its ability to give a useful summary of the data is in line with the model development situation. Being clear about one's intentions at the outset of the analysis is thus crucial. In sum, it is relatively easy to determine that a model is a bad fit to the data; it is a far more difficult and nuanced task to determine that a model is a good fit to the data and supportive of a hypotheses implied by the model.

Special topics

A brief overview of several special topics is provided so that readers are familiar with the concepts and know where to look for additional information.

Categorical measured variables

Outcomes in nursing research are frequently not distributed on a continuous scale (e.g., survival status, presence or absence of a symptom). Link functions such as the probit and logistic functions are used to model categorical *endogenous* variables (Agresti, 1990; Hosmer & Lemeshow, 2000; Muthén, 1983, 1984). Alternative estimation procedures are also needed if indicators in CFA are not measured continuously (Muthén, 1983, 1984), including when Likert-type ratings (with fewer than four response options, West et al., 2012) are used in self-report questionnaires. To ensure that expected statistical properties hold, alternative approaches should be used for estimation (Flora & Curran, 2004; Wirth & Edwards, 2007). A weighted least squares (WLS) estimator, or a robust version for use with smaller sample size, is often used for CFA with categorical or ordinal data (Lei & Wu, 2012; Muthén, du Toit, & Spisic, 1997). Analytic relationships with some IRT models also suggest that response structures can be analyzed in an IRT environment (Baldwin & Wainer, this volume; Thissen & Steinberg, 1986).

Nested data

Clustered datasets comprise individuals nested within groups (e.g., patients nested within providers) and occur frequently in nursing research. Ignoring clustering in the analysis of such data can result in standard errors that are too small and inflated type I error rates (Bland, 2004; Grittner &

Lahmann, this volume). Several SEM packages now include estimation procedures suitable for such multilevel data, either by modeling the data in a multilevel framework where “between” and “within” individual data are explicitly modeled (Muthén, 1994; Preacher, Zyphur, & Zhang, 2010), or by using estimators that correct the chi-square and standard errors for non-independence (Muthén & Satorra, 1995).

Longitudinal growth curve modeling

Data measured longitudinally are a special case of multilevel data with time points nested within individuals (Duncan, Duncan, Strycker, Li, & Alpert, 1999). Longitudinal designs are common in nursing research (e.g., changes in symptoms over time; response to treatment over time). Latent growth curve models for longitudinal models can be estimated in an SEM framework (Singer & Willett, 2003, Chapter 8).

Multiple group models

SEM models can be estimated simultaneously in one or more groups to determine whether some or all of the parameter estimates differ significantly across groups. The multiple group approach answers questions about statistical moderation (i.e., whether the same model is operative across subgroups of individuals; Aiken, Stein, & Bentler, 1994) and measurement invariance (i.e., whether a construct is being measured in an equivalent way across groups (e.g., Sousa et al., 2012).

Mean structures

The models discussed in this chapter focused on relationships in the variance-covariance matrix, which summarizes mean-centered data. Means can also be modeled (Browne & Arminger, 1995) and are essential to understanding change over time in growth curve modeling, and comparing means among groups in multiple group analysis.

Latent variable mixture modeling

Mixture modeling refers to a family of models that classify individuals into subpopulations based on heterogeneity in the data (Muthén, 2002; Stoddard, this volume). Class membership reflects similarity of individuals along a dimension of interest (e.g., behavior, attitudes) and dissimilarity from those in other classes. Mixture modeling determines whether a *categorical* latent variable underlies the measured variables, where the category represents an individual projected class assignment. Schmiede, Meek, Bryan, and Petersen (2012) provided an overview of the use of these techniques in nursing research.

Sample size

Estimating required sample size for SEM is not straightforward and cannot be accomplished with traditional power analysis software packages. Findings from the robustness literature may provide useful guidance about sample size (e.g., Curran, West, & Finch, 1996; Gagné & Hancock, 2006). We refer readers to MacCallum, Browne, and Sugawara (1995), Muthén and Muthén (2002), and Satorra and Saris (1985) for three approaches to power analysis and to Thoemmes, MacKinnon, and Reiser (2010) for a nice illustration of Muthén and Muthén’s (2002) Monte Carlo procedure for determining sample size.

SEM software

The first software package for SEM – LISREL – was distributed in 1974. Since that time, other SEM packages have been introduced and updated, such that users have an excellent array of programs from which to choose. Selection of a program is often based on personal preference; for example, having taken a course where a particular package was used and getting accustomed to its language and specifications. Making decisions based on personal preference is justified for basic SEM applications because major SEM packages are nearly equivalent for these uses (e.g., CFA, path analysis). Differences appear primarily when more sophisticated analyses (e.g., multi-level modeling, multiple groups analysis, mixture modeling) are undertaken. We briefly review the specifications, strengths, and potential shortcomings of the four commonly used and widely available commercial SEM packages in their most current versions: LISREL 9.1, AMOS 18.0, EQS 6.2, and *Mplus* 7.11 (see Table 8.2).

A defining feature of LISREL has been its reliance in both input and output on matrix notation using Greek symbols. This necessitates understanding the relationship between matrix notation and schematics that underlie SEM models. LISREL8 introduced a command language requiring that users specify only the observed and latent variables and hypothesized regression paths. Knowing how to write matrix-based LISREL code is thus no longer necessary, but it is still useful to have some knowledge of the notation because it is a standard and easily understood representation of SEM models and is useful for interpreting output of other programs (e.g., *Mplus* technical output follows LISREL notation). (We have followed LISREL notation in this chapter.)

AMOS 18.0 is well-known for its graphical interface that allows users to “draw” model specifications. The user-friendliness of the input specification technology is clearly a strength, but some also worry that AMOS is often the choice of users who are unfamiliar with the underlying assumptions of SEM (Byrne, 2012).

EQS 6.2 utilizes an equation-based text window for model specification. This equation-based structure is intuitive for those familiar with SEM as the simultaneous estimation of linear regression equations. EQS is not able to estimate mixture models, though Bentler (2006) asserted that this capability will be added once there is – in his view – better clarity on the best way to estimate these models.

Mplus now has both text-based command language and a graphical interface for model specification. The number of commands necessary to specify even fairly complex models is quite small, making *Mplus* relatively easy to program, although we recommend that users write code for each aspect of the model rather than relying on default settings. *Mplus* is in a class by itself in terms of its flexibility to estimate a wide range of complex models with various types of data (e.g., non-normal, categorical, count, censored).

In addition to these common commercial SEM software packages, R software includes several SEM options (Boker et al., 2011; Fox, 2006; Rosseel, 2012). R software is a free, open-source software environment for statistical computing. Two approaches to carrying out SEM in R include connecting R with external SEM software packages (e.g., *Mplus*, EQS) or using an R package dedicated to SEM (e.g., *lavaan*, *openMx*, *sem*).

Additional resources

This chapter was intended to serve as a primer on latent variable modeling and SEM for nursing research. For more information on these topics, a variety of comprehensive books and edited volumes are available (e.g., Bollen, 1989; Byrne, 1998; Hayduk, 1987; Hoyle, 2012; Kline, 2005; Loehlin, 1992; Marcoulides & Schumacker, 1996). The journal *Structural Equation Modeling* is

<i>Feature</i>	<i>LISREL 9.1</i>	<i>EQS 6.2</i>	<i>AMOS 18.0</i>	<i>Mplus 7.11</i>
Import Data Types	Raw data ^a Preprocess with PRELIS	Raw data ^a Covariance matrix SPSS.sav files	MS Excel MS Access SPSS.sav files	Raw data ^a Covariance matrix
Specification Options	Syntax (LISREL commands) Interactive (SIMPLIS) Graphical interface	Syntax (equation-based) Interactive (dialogue boxes) Graphical interface	Graphical interface Text editor (equation-based)	Syntax (<i>Mplus</i>) Interactive language generator Graphical interface
Estimation Methods	ULS, GLS, ML, ML robust, WLS, DWLS	ULS, GLS, ML, ML robust, ADF	ULS, GLS, ML, ML robust, ADF, SLS	ULS, GLS, WLS, ML, ML robust (MLM, MLM, MLR), ULSMV, WLSM, WLSMV
Fit Indices	Many	Many	Many	Fewer
Missing Data	MI, FIML, MCMC, EM	Imputation during data management based on means, group means, regression, EM; ML	FIML	ML Special feature for estimating over multiple imputed datasets
Robust ML	Yes	Yes	Yes	Yes
Latent Class/Mixture Models	Yes	No	Yes	Yes
Multilevel Analysis	Yes	Yes	Yes	Yes
Strengths	Matrix notation facilitates understanding of SEM methodology	Particularly user friendly for multiple group models; comprehensive data management; exploratory missing data analyses; many methods for non-normal data	Specification search function ^b	User friendly for all types of growth modeling, mixture modeling; Especially good at handling categorical, count, and censored endogenous variables; Most flexible of all programs in types of models estimated
Information	www.ssicentral.com/lisrel	www.mvsoft.com	www-01.ibm.com/software/analytics/spss/products/statistics/amos/	www.statmodel.com

Note: ML = maximum likelihood; MI = multiple imputation; FIML = full information maximum likelihood; EM = expectation maximization; ULS = unweighted least squares; GLS = generalized least squares; WLS = weighted least squares; DWLS = diagonally weighted least squares; ADF = asymptotic distribution free; SLS = scale-free least squares; MCMC = Markov chain monte carlo.

^aRaw data are data in text formats.

^bThe specification search function is essentially an exploratory approach to structural relationships among variables – which could also be viewed as a weakness (Byrne, 2012).

exclusively devoted to SEM. This journal includes teaching articles that provide a comprehensive and easy to understand discussion around a specific topic in each issue. Finally, the SEMNET listserv (www2.gsu.edu/~mkteer/semnet.html) is a valuable resource for investigators wishing to connect with or to ask specific questions of those with SEM expertise.

References

- Agresti, A. (1990). *Categorical data analysis*. New York, NY: Wiley.
- Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural analyses of clinical sub-population differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology*, 62, 488–499.
- Ajzen, I. & Madden, T. J. (1986). Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control. *Journal of Experimental Social Psychology*, 22, 453–474.
- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317–332.
- Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103, 411–423.
- Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2013). Causality and endogeneity: Problems and solutions. In D. V. Day (Ed.), *The Oxford handbook of leadership and organizations*. Oxford: Oxford University Press.
- Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.
- Bentler, P. M. (2006). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–606.
- Bland, J. M. (2004). Cluster randomised trials in the medical literature: Two bibliometric surveys. *BMC Medical Research Methodology*, 4(21) [Electronic Resource]. Retrieved from www.biomedcentral.com/1471-2288/4/21
- Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., . . . Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, 76, 306–317.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Box, G. E. P. (1979). Some problems of statistics of everyday life. *Journal of the American Statistical Association*, 74, 1–4.
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62–83.
- Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean and covariance structure models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). New York, NY: Plenum.
- Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.
- Byrne, B. M. (2012). Choosing structural equation modeling computer software: Snapshots of LISREL, EQS, Amos, and Mplus. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 307–304). New York, NY: Guilford Press.
- Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, 41, 189–225.
- Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, 105, 317–327.
- Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the “problem” of sample size: A clarification. *Psychological Bulletin*, 109, 512–519.
- Cudeck, R., Klebe, K. J., & Henly, S. J. (1993). A simple Gauss-Newton procedure for covariance structure analysis with high-level computer languages. *Psychometrika*, 58, 211–232.
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1, 16–29.
- Duncan, T. E., Duncan, S. C., Strycker, L. A., Li, F., & Alpert, A. (1999). *An introduction to latent variable growth curve modeling*. Mahwah, NJ: Lawrence Erlbaum.

- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466–491.
- Folse, V. N. (2007). The Family Experience with Eating Disorders Scale: Psychometric analysis. *Archives of Pediatric Nursing*, 21, 210–221.
- Fox, J. (2006). Structural equation modeling with the sem package in R. *Structural Equation Modeling*, 13, 465–486.
- Fox, J. (2009). *A mathematical primer for social statistics*. Thousand Oaks, CA: Sage.
- Gagné, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, 41, 65–83.
- Gibbons, R. D., & Hedecker, D. (1992). Fill-information bi-factor analysis. *Psychometrika*, 57, 423–436.
- Grayson, D., & Marsh, H. W. (1994). Identification with deficient rank loading in confirmatory factor analysis: Multitrait-multimethod matrices. *Psychometrika*, 59, 121–134.
- Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Baltimore, MD: Johns Hopkins University Press.
- Hayduk, L. A., & Glaser, D. N. (2000). Jiving the four-step, waltzing around factor analysis, and other serious fun. *Structural Equation Modeling*, 7, 1–35.
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research*, 26, 329–367.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: Wiley.
- Hoyle, R. H. (Ed.). (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage.
- Hoyle, R. H. (Ed.). (2012). *Handbook of structural equation modeling*. New York, NY: Guilford Press.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109–133.
- Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, 43, 443–477.
- Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long, *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.
- Jöreskog, K. G., & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer software]. Skokie, IL: Scientific Software.
- Kaplan, D., Harik, P., & Hotchkiss, L. (2001). Cross-sectional estimation of dynamic structural equation models in disequilibrium. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 315–340). Lincolnwood, IL: Scientific Software.
- Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112, 165–172.
- Kenny, D. A., & Milan, S. (2012). Identification: A nontechnical discussion of a technical issue. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 145–163). New York, NY: Guilford.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford.
- Lei, P. W., & Wu, Q. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 164–180). New York, NY: Guilford.
- Loehlin, J. C. (1992). *Latent variable models: An introduction to factor, path, and structural analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Lynn, M. R., Morgan, J. C., & Moore, K. A. (2009). Development and testing of the satisfaction in nursing scale. *Nursing Research*, 58, 166–174.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1995). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130–149.
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111, 490–504.
- MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, 114, 185–199.
- Marcoulides, G. A., & Schumacker, R. E. (Eds.). (1996). *Advanced structural equation modeling: Issues and techniques*. Mahwah, NJ: Lawrence Erlbaum.
- Marsh, H. W., Hau, K. T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A Festschrift for Roderick P. McDonald* (pp. 225–340). Mahwah, NJ: Lawrence Erlbaum.

- Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11, 320–341.
- McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7, 64–82.
- Melnnyk, B. M., Crean, H. F., Feinstein, N. F., & Fairbanks, E. (2008). Maternal anxiety and depression after a premature infant's discharge from the neonatal intensive care unit. *Nursing Research*, 57, 383–394.
- Micceri, T. (1989). The unicorn, the normal distribution, and other improbable creatures. *Psychological Bulletin*, 105, 156–166.
- Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, 22, 43–65.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115–132.
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22, 376–398.
- Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, 29, 81–117.
- Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Retrieved from http://pages.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf
- Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25, 267–316.
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 4, 599–620.
- Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91). New York, NY: Guilford Press.
- Phillips, L. J., & Stuifbergen, A. K. (2009). Structural equation modeling of disability in women with fibromyalgia or multiple sclerosis. *Western Journal of Nursing Research*, 31, 89–109.
- Preacher, K., Zyphur, M., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15, 209–233.
- Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114, 552–566.
- Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, 23, 51–67.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36.
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.
- Satorra, A. & Saris, W. (1985). The power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, 50, 83–90.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Schmiege, S. J., Meek, P., Bryan, A. D., & Petersen, H. (2012). Latent variable mixture modeling: A flexible statistical approach for identifying and classifying heterogeneity. *Nursing Research*, 61, 204–212.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.
- Skrondal, A., & Rabe-Hesketh, S. (2004). Identification and equivalence. In *Generalized latent variable modeling. Multilevel, longitudinal, and structural equation models* (pp. 135–158). Boca Raton, FL: Chapman & Hall/CRC.
- Sousa, K. H., Ryu, E., Kwok, O. M., Cook, S. W., & West, S. G. (2007). Development of a model to measure symptom status in persons living with rheumatoid arthritis. *Nursing Research*, 56, 434–440.
- Sousa, K. H., West, S. G., Moser, S. E., Harris, J. A., & Cook, S. W. (2012). Establishing measurement invariance: English and Spanish Pediatric Asthma Quality of Life Questionnaire. *Nursing Research*, 61, 171–180.
- Spence Laschinger, H. K., & Leiter, M. P. (2006). The impact of nursing work environments on patient safety outcomes: The mediating role of burnout/engagement. *Journal of Nursing Administration*, 36, 259–267.
- Stuart, A. S., & Ord, J. K. (1991). Estimation. In *Kendall's advanced theory of statistics* (5th ed.) (pp. 605–648). New York, NY: Oxford University Press.

- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567–577.
- Thoemmes, F., MacKinnon, D. P., & Reiser, M. R. (2010). Power analysis for complex mediational designs using Monte Carlo methods. *Structural Equation Modeling*, 17, 510–534.
- West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford.
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12, 58–79.
- Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 14, 199–213.
- Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology* 2000, 30, 167–202.

9

GENERAL GROWTH MIXTURE MODELS

Sarah A. Stoddard

Health is complex and dynamic, influenced by both individual characteristics (genetics, biology, behavior) and an individual's interaction with their social and physical environment over the course of a lifespan (Henly, Wyman, & Findorff, 2011). Understanding the factors (and timing of those factors) that influence health over time provides a basis for identifying individuals at risk for adverse trajectories or outcomes, and offers insight on the timing and effects of interventions to remediate compromised health. Equally important is the ability to identify and explain variation in health behaviors, health outcomes, and responses to interventions over time. Identification of population subgroups whose members' health changes similarly over time provides the opportunity for tailoring interventions to specific individuals in those subgroups. In intervention research, assessment of health status over time permits characterization of response (quick or slow, temporary or durable, constant, or changing rate of change) and identification of responders and nonresponders (e.g., Muthén, Brown et al., 2002; Nies & Sun, 2008; Stein, Dickstein, Schuster, Litz, & Resick, 2012; Wills et al., 2012).

Conventional latent growth curve models (LGCMs) estimate change parameters for each individual (random effects, microparameters) and variance parameters around average growth parameters (fixed effects) in a single population (Singer & Willett, 2003). Growth mixture modeling was first introduced to understand variation in growth patterns over time (change patterns) within a population attributable to latent subgroups – sometimes called “latent classes” (Muthén & Shedden, 1999; Verbeke & Lesaffre, 1996). Thus, the general growth mixture model (GGMM) is an extension of the LGCM that identifies distinct subgroups of cases with similar growth trajectories, and allows individual trajectories to vary around class-specific mean trajectories. The GGMM relaxes the assumption that all individuals are drawn from a single population with common fixed effects by including latent trajectory classes (subpopulations) with separate growth parameters and variance components. Another related approach, group-based trajectory modeling (Nagin & Odgers, 2010), occupies a middle position, positing the existence of latent subgroups but with no within-class variation. The key assumption in both GGMM and group-based trajectory modeling is that growth parameters arise from a mixture of finite populations (cf., Titterton, Smith, & Makov, 1985), sometimes referred to as “population heterogeneity.”

The purpose of this chapter is to explore how to utilize general growth mixture modeling to assess similarities in interindividual differences in intraindividual change in health behavior or

outcomes over time. It will describe and explain the methodology behind GGMM and use an example to demonstrate the modeling process step by step.

The data example

The illustration uses data from a study that examined trajectories of hopelessness in a sample of adolescents living in impoverished urban neighborhoods (Stoddard, Henly, Sieving, & Bolland, 2011). The sample included 723 African-American youth (49% female) living in impoverished neighborhoods in the southern US who participated in an annual, community-based survey from 1998 through 2006 (Bolland, Bryant et al., 2007; Bolland, Lian, & Formichella, 2005).

The purpose was to examine the relationships among social connections, hopelessness trajectories, and subsequent violent behavior across the adolescent years (Figure 9.1) using a GGMM that integrated theory, temporal design, and statistical model (following Collins, 2006). It was hypothesized that (a) hopelessness would change during middle adolescence due to cognitive changes and identity formation, and that distinctly different patterns of change – trajectory classes – would emerge; (b) known groups [gender] and variation in social connections and violent behavior during early adolescence would predict hopelessness trajectory class membership and within-class growth parameters; and (c) trajectory classes characterized by increasing hopelessness would be associated with a higher probability of violence with a weapon during late adolescence.

The trajectory variable was hopelessness, measured from one to four times across middle adolescence (ages 13–16) using the six-item Brief Hopelessness Scale (BHS; Bolland, McCalum, Lian, Bailey, & Rowan, 2001). Age was measured in years at the time each survey was filled out and centered on age 13 for the trajectory analysis. Covariates measured once during

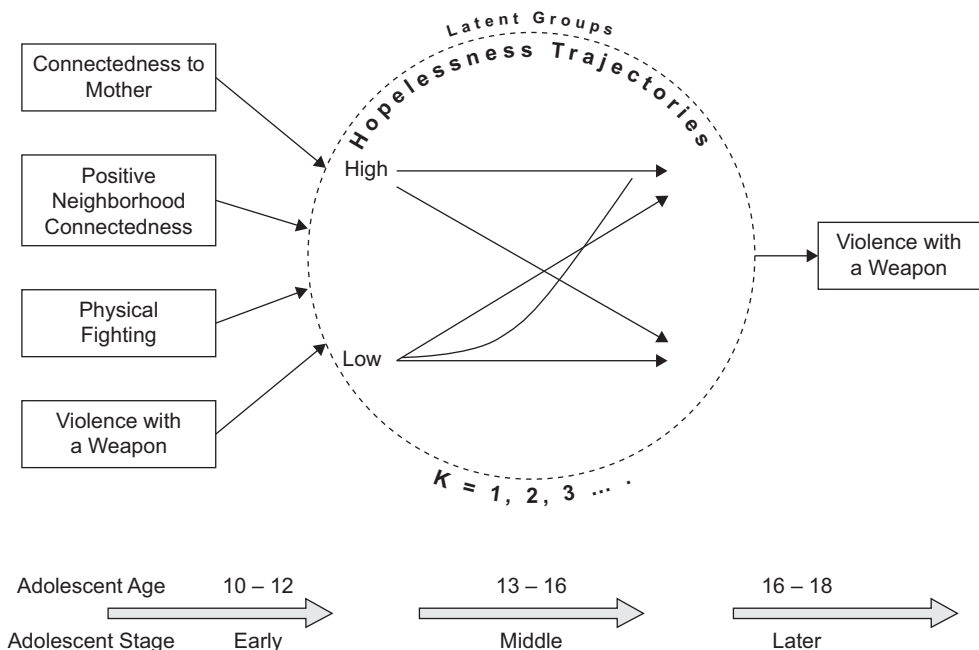


Figure 9.1 General growth mixture models: theoretical hope/hopelessness trajectory model. From Stoddard, Henly, Sieving, and Bolland (2011). Used with permission.

early adolescence (approximately age 12) were: connection to mother (score range 0–6), positive neighborhood connectedness (score range 0–6), and physical fighting and violence with a weapon during early adolescence. Physical fighting was assessed with a single item: *In the past three months (90 days), were you in a physical fight?* Response options were coded so that 0 = *No fighting* and 1 = *Yes/any level of fighting*. Violence with a weapon was dichotomous variable scored 0 = *none* and 1 = *threatening someone with a weapon in the past 90 days, pulling a weapon on someone in the past 90 days, shooting someone in the past year, or stabbing someone in the past year*. The distal outcome – measured at approximately age 16 – was violence with a weapon during later adolescence. The same scheme for scoring violence with a weapon was used in early and later adolescence. Gender was self-reported (0 = *girl* and 1 = *boy*). Gender was included as a known group in the analysis because previous research on relationships among social connectedness, the dynamics of hopelessness, and violence during adolescence suggested them to be different for boys and girls (Farrington, Langan, & Tonry, 2004; Loeber, & Stouthamer-Loeber, 1998; Tremblay et al., 2004).

General growth mixture modeling

The general growth mixture model (GGMM) (Li, Duncan, Duncan, & Acock, 2001; Muthén, 2001; Muthén & Shedden, 1999) is a model for means and covariances that incorporates latent growth curve analysis and a mixture of distinct latent classes whose members are characterized by similar change (intraindividual variation) in a health outcome over time. The within-class pattern of change may be influenced by prior covariates, as well as moderated by membership in known groups, and may be predictive of distal outcomes that can be dichotomous, multinomial, ordered or continuously varying in nature. Inclusion of covariates and distal outcomes helps to explicate differences in change that define the latent classes (Muthén & Shedden, 1999, p. 464), thereby allowing for greater understanding of the differing response patterns. Inclusion of known groups in the model, such as those defined by gender or assignment to experimental condition, reflects a moderation hypothesis whereby model parameters potentially vary as a function of membership in the known group.

Using this GGMM approach, the aim of the analysis used in Stoddard et al. (2011), and illustrated here, was to estimate a model that characterized individual hopelessness trajectories during middle adolescence; identified latent trajectory classes; used covariates, known groups, and distal outcomes to define trajectory classes; and linked class membership with the distal outcome of violence with a weapon during later adolescence. Building a GGMM is a complex endeavor. Each step in the development of the model will be discussed: (a) the preliminary exploratory data analysis; (b) modeling individual latent growth curves; (c) latent class modeling with covariates, known groups, and distal outcomes; (d) model estimation; and (e) model testing and selection.

Preliminary analysis: Explore individual patterns of change

A preliminary, exploratory analysis is conducted to examine patterns of change in the trajectory variable for the cases in the sample. This descriptive analysis is important as it can reveal the nature of each person's temporal pattern of growth or decline (each individual's trajectory), as well as differences in pattern of change among the individuals in a sample. In our example, hopelessness is the trajectory variable.

To explore individual trajectories, empirical growth plots should be created to provide graphic description of how different individuals change over time (e.g., Singer & Willett, 2003, Chapter 2). Nonparametric smoothing can be used to summarize each individual's pattern of change without committing to a specific functional form. The individual trajectories should be examined with particular attention to the intercepts, direction of change, and constancy of change.

Similarities and differences among individuals in the overall pattern of change (e.g., linear versus nonlinear) should be assessed. Ordinary least-squares (OLS) regression can be used to obtain preliminary estimates of coefficients for linear change (i.e., intercepts and slopes for each person). Visual impressions from the graphs and information about the distribution of intercepts, change coefficients, and explained and residual variances can aid in identifying reasonable functional forms for overall change trajectories. The number and temporal spacing of individuals' data points and theoretical perspectives will also influence the functional forms to be considered.

Modeling individual latent growth curves

A wide variety of change trajectories may be conveniently described by linear and quadratic functions (Henly et al., 2011; Singer & Willett, 2003). A linear function implies that there is a constant rate of change over the defined time period. In the case of the example data, a model with a linear growth function assumes that hopelessness has a constant rate of change throughout middle adolescence, whereas a quadratic function assumes that hopelessness may either accelerate or decelerate.

The quadratic growth (change) function for variable y for individual i measured at time t can be formally stated as

$$y_{it} = \eta_{i0} + \eta_{i1}t + \eta_{i2}t^2 + \varepsilon_{it} \quad (9.1)$$

where η_{i0} , η_{i1} , and η_{i2} represent the intercept, slope, and quadratic terms of the growth function for person i and ε_{it} is the residual at the time points. (When $\eta_{i2} = 0$, the change function is linear.) More generally, the following equation in matrix notation defines the conventional growth model for a continuous observed trajectory variable y with m continuous growth parameters for individual i across p theoretically interesting time points:

$$\mathbf{y}_i = \mathbf{\Lambda}_y \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \quad (9.2)$$

where in our example, $\mathbf{y}_i$ is a column vector containing observed hopelessness scores for person i on p occasions that may vary from one individual to the next; $\boldsymbol{\eta}_i$ contains m growth parameters (coefficients for intercept, linear and/or quadratic terms); $\mathbf{\Lambda}_y$ is a $(p \times m)$ fixed design matrix for temporal effects; and $\boldsymbol{\varepsilon}_i$ contains residuals for person i (deviations of the observed scores around the model expectation). The residuals $\boldsymbol{\varepsilon}_i$ are assumed to have a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Theta}$.

Equation 9.2 can be rewritten explicitly to show the elements for a specific design. For example, when $p = 4$ measurements (say at times $t = 0, 1, 2$, and 3) are obtained on trajectory variable y for person i with individual growth parameters η_{i0} , η_{i1} , and η_{i2} , the expression for a quadratic growth model is:

$$\begin{bmatrix} y_{i0} \\ y_{i1} \\ y_{i2} \\ y_{i3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} \eta_{i0} \\ \eta_{i1} \\ \eta_{i2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i0} \\ \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \end{bmatrix}$$

$$\mathbf{y}_i = \mathbf{\Lambda}_y \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$

The first column of the design matrix $\mathbf{\Lambda}_y$ contains values 1 that are indicators used to include the intercept term η_{i0} in the model. The second column entries are the values of t at each of the

p observations. The third column entries are the t^2 values associated with each observation time that allow the quadratic effect to be estimated.

Adding covariates, distal outcomes, and known groups

GGMM attributes some variation in individuals' growth curves to their membership in distinct latent classes that have different mean growth curves. GGMM estimates the mean growth curve for each potential class via mean growth parameters (intercepts, slopes, etc.) that differ for each class (Li, Duncan, Duncan, & Acock, 2001; Muthén, 2004). Under the model, individual growth parameters $\boldsymbol{\eta}_i$ can be linked with latent classes using

$$\boldsymbol{\eta}_i = \mathbf{A}\mathbf{c}_i + \boldsymbol{\zeta}_i, \quad (9.3)$$

where $\boldsymbol{\eta}_i$ is a vector of individual growth parameters as shown above in Equation 9.2. $\mathbf{A}$ is a matrix containing latent growth parameter means for each latent class k (one column for each class), and the K -dimensional vector $\mathbf{c}_i$ is an indicator vector with elements equal to 0 when i is not in a latent class and equal to 1 when i is in a particular class. Thus, the matrix product $\mathbf{A}\mathbf{c}_i$ assigns to person i the vector of mean growth parameter values for the class in which person i is a member. The vector $\boldsymbol{\zeta}_i$ contains deviations of growth parameters for individual i from the growth parameter mean vector for their class. The deviations $\boldsymbol{\zeta}_i$ are assumed to have a multivariate, normal distribution with mean $\mathbf{0}$ and variance-covariance matrix equal to $\boldsymbol{\Psi}$. Under this model, individual growth parameters are a composite of their latent class average parameters and deviations of the individual from the class average.

Covariates

Covariates serve two purposes in GGMM. First, an important part of GGMM is the prediction of class membership probabilities from covariates. In GGMM, class membership can be better defined and explained (predicted) by the incorporation of covariates. Models estimating class membership without covariates (using only trajectory variables) may be less able to accurately differentiate classes and may lack important support provided by both theory and covariates (Muthén, 2004). Second, covariates may be used to explain within-class variability in the individual latent trajectory parameters, and the investigator can explore whether the effects of the covariates vary across classes.

Because classes in GGMM are unordered categories, multinomial logistic regression for unordered polytomous variables can be used to relate the predictors (covariates) to the individual probabilities of latent trajectory class membership (Muthén & Shedden, 1999; Bauer & Curran, 2003). For a model with K latent classes, define the K -dimensional vector $\mathbf{c}_i$ to be an indicator vector with elements c_{ik} equal to 0 when person i is not in a latent class k , and equal to 1 when i is in a particular class k . Let $\boldsymbol{\pi}_i$ be the vector of probabilities of membership in each latent class k ($k = 1, 2, \dots, K$) for person i , conditioned on their covariate score vector $\mathbf{x}_i$; with element $\pi_{ik} = p(c_{ik} = 1 | \mathbf{x}_i)$. Next, let the $(K - 1)$ -dimensioned vector of log odds of membership in class k ($k = 1, 2, \dots, K - 1$) with respect to membership in class K (the reference class) be $\text{logit}(\boldsymbol{\pi}_{ik})$, with element $\text{logit}(\pi_{ik}) = \log[\pi_{ik} / \pi_{iK}]$. Then

$$\text{logit}(\boldsymbol{\pi}_i) = \boldsymbol{\alpha}_c + \boldsymbol{\Gamma}_c \mathbf{x}_i, \quad (9.4)$$

where $\boldsymbol{\alpha}_c$ is a $(K - 1)$ -dimensional vector of mean log odds for classes $k = 1, 2, \dots, K - 1$ and $\boldsymbol{\Gamma}_c$ is dimensioned $(K - 1)$ by q , and contains coefficients linking individual covariate scores to changes

in log odds of class membership. The second term $\mathbf{\Gamma}_c \mathbf{x}_i$ adjusts overall log odds of trajectory class membership α_c for scores on covariates included in the model.

As mentioned above, covariates can also be used to explain within-class variability in the individual latent trajectory parameters. Under the model, individual growth parameters η_i can then be linked with q covariates $\mathbf{x}_i$ and k latent classes using

$$\eta_i = \mathbf{A} \mathbf{c}_i + \mathbf{\Gamma}_\eta \mathbf{x}_i + \zeta_i. \quad (9.5)$$

Equation 9.5 is an extension of Equation 9.3 which adds $\mathbf{\Gamma}_\eta$, an $m \times q$ matrix linking q covariates with m growth parameters, and $\mathbf{x}_i$, the vector of scores on the covariates for person i . The middle term of Equation 9.5, $\mathbf{\Gamma}_\eta \mathbf{x}_i$, reflects an adjustment to the class-specific mean growth parameters due to the covariates.

The model represented by Equation 9.5 can be described as an equal prediction model because it assumes that the effect of covariates on growth parameters is equivalent across latent classes. However, the model can be expanded to include class-specific elements in the matrix $\mathbf{\Gamma}_\eta$ that describe the impact of covariates on individual growth parameters for a given class. In practical terms, the equal prediction model is simpler, and therefore easier to estimate and interpret. On the other hand, the group-specific prediction model will manifest better fit to the sample data because it is more highly parameterized – particularly when the number of latent classes is large. In considering whether to use equal prediction or group-specific prediction models for the covariates, one should ask, *Is there any reason to believe that covariates differentially impact change, depending on how i (the individual) is changing?*

Distal outcome

Distal outcomes may also be included in the model. Distal outcomes are subsequent outcomes that may be predicted from growth latent class membership and covariates. In our example, violence with a weapon during later adolescence is the binary distal outcome. GGMM can also include continuous and multiple category distal outcome variables. GGMM gives the probability of a dichotomous outcome, like violence with a weapon during later adolescence (u) for person i in latent class $c_i = k$ and with covariates $\mathbf{x}_i$ as

$$\tau_i = P(u_i = 1 | c_i = k, \mathbf{x}_i) \quad (9.6)$$

The related log odds of $P(u_i = 1 | c_i = k, \mathbf{x}_i)$ versus $P(u_i = 0 | c_i = k, \mathbf{x}_i)$ is $\log[\tau_i / (1 - \tau_i)]$, which is modeled as:

$$\text{logit}(\tau_i) = \Lambda_u \mathbf{c}_i + \mathbf{K}_u \mathbf{x}_i \quad (9.7)$$

The vector $\Lambda_u(1 \times K)$ contains the conditional log odds for outcome u for each latent growth class k , and $\mathbf{K}_u(1 \times q)$ is a vector of regression coefficients of log odds for u on the covariates $\mathbf{x}_i$ (measured during early adolescence in our example). The model can be extended to allow the regression of the log odds for u on the $\mathbf{x}_i$ to vary by class. Thus, for our example, the log odds associated with later adolescent violence with a weapon is a linear function of the log odds specific to each hopelessness change latent class during middle adolescence and covariates measured during early adolescence.

Known groups

In our example, gender is included as a known group that is observed and functions as a moderator variable. In this case, all model parameters potentially vary as a function of membership in the observed group. Equation 9.6 is changed to include the identification of gender groups j :

$$\tau_{ij} = P(u_{ij} = 1 | c_{ij} = k, \mathbf{x}_i) \quad (9.8)$$

Treating gender as a moderator is often preferred to treating gender as a covariate in the models. Gender as a moderator (or known group) allows the class probabilities to vary across the observed group in the sample. Latent classes are partially known, which means that, for example, girls could be in only those latent classes derived from mixture models carried out for girls. Including gender as a covariate would impose equalities between genders that may not be valid (such as structure of classes, within-class variability, effects of covariates, and associations between class membership and distal outcomes).

Estimation

In *Mplus*, the maximum likelihood (ML) estimator for GGMM is based on the expectation-maximization (EM) algorithm (Muthén, 2001; Singer & Willett, 2003). The numerical, iterative solution is obtained as follows: The conditional probability of individual i belonging to the latent class K or the posterior probabilities of class membership, and an individual's score on the growth factors is estimated. For each individual in the sample, these probabilities estimate the probability of the individual's belonging to each trajectory class. The EM algorithm starts with a random split of people into classes. Reclassification is based on an improvement criterion, with reclassification occurring until the best classification of people is found. Multiple random start values are used to avoid solutions at a local minimum (Muthén & Muthén, 1998–2007). The solution is the set of parameters associated with the best log likelihood.

Model evaluation and selection

To evaluate model quality, a combination of statistical fit, theoretical plausibility, and usefulness criteria should be used to identify the best model, including the most appropriate number of classes. *How well does the model account for the data?* (i.e., statistical fit) and *Does the model make sense?* (i.e., theoretical or conceptual plausibility) should both be addressed and planned uses of the model should be considered (Browne, 1982; Cudeck & Henly, 2003; Henly, Vermeersch & Duckett, 1998). Goodness of fit, parameter estimates, standard errors, z - or t -statistics and p -values, and residuals should be examined.

Overall fit

Candidate models should be compared for overall fit using the Akaike information criterion (AIC), Bayesian information criterion (BIC), and the adjusted BIC. All three criteria are based on the negative log likelihood of the model with a penalty function for the number of estimated parameters. Comparatively smaller values for all fit indices indicate a better fitting model (Li et al., 2001; Muthén, 1998–2004; also see Schmiede & Bryan, this volume).

Entropy

Usefulness can be examined by considering classification quality or, *How well do the members fit in each class?* Entropy is a summary measure of how well people are classified into latent classes; i.e., that cases within classes are like each other and not like those in other groups in the way they change. Entropy is measured on a 0 to 1 scale, with values closer to 1 being more desirable (Muthén, 2004). Values $\geq .90$ are suggested to indicate good classification quality (Acock, 2009). The number of participants in a latent class is also a consideration; classes with very few members are not likely to be interpretable or replicate.

Missing data (missingness)

When modeling change over time, it is assumed that each person's observed records are a random sample of data from his or her underlying true growth trajectory. Missingness mechanisms must be considered for the serial data, which may be missing completely at random (MCAR), missing at random (MAR), or nonignorable missingness (Raudenbush, 2001). MCAR means that the probability of a missing response depends on neither the observed nor nonobserved response. MAR occurs when the probability of missing a time point is independent of the missing data given the observed data. The key question for MAR is whether it is safe to assume that the probability of missingness is unrelated to unobserved concurrent outcomes (Singer & Willett, 2003). This assumption is reasonable when the observed data captures key confounding influences, like variables that predict both attrition and the outcome of interest. Under MAR, estimation is unbiased if all the data are used in the analysis and a fully efficient estimation procedure is used (like ML estimation of multilevel mixed models) (Raudenbush, 2001). ML efficiently uses all available time points to estimate the model, assuring unbiased estimation. Under these conditions, the mechanism that produces the missingness is considered ignorable. Nonignorable missingness occurs when data are neither MCAR or MAR, and occurs in longitudinal studies when missing data patterns are intermittent or "monotone missing" (dropouts who do not complete a measurement protocol) (Tseng, Elashoff, Li, & Li, 2012).

Data example: GGMM for hopelessness in adolescents

The scatterplots of hopelessness measurements by age during middle adolescence (Figure 9.2) for subsamples of girls and boys suggest that a variety of functional forms for change may be appropriate (e.g., linear, quadratic). OLS regression was then used to explore hopelessness trajectories for linear change functions (Figure 9.3; Table 9.1).

The average estimated intercept for hopelessness among girls was 1.24 and average estimated slope was $-.07$, indicating that girls on average reported low initial levels of hopelessness at age 13 that then decreased slightly each year. The average estimated intercept for hopelessness among boys was higher (1.62) and, on average, boys' levels of hopelessness showed very little change across the middle adolescent years (average slope = $.03$). Hopelessness scores for individual boys and girls were scattered widely around their gender averages, showing considerable variability both in their fitted initial status, or intercept ($SD = 2.00$ and $SD = 1.80$, respectively), and fitted rate of change, or slope ($SD = .67$ and $SD = .66$, respectively). The slope reflects the coding of time and, in our example, refers to change in hopelessness per year. The observed correlation between intercept and slope was $-.74$ for boys and $-.81$ for girls. Girls with higher initial hopelessness tended to decrease more over time than boys with an equivalent level of initial hopelessness. Taken together, the linear change function seemed reasonable, but accelerating change (quadratic function) appeared to characterize some trajectories.

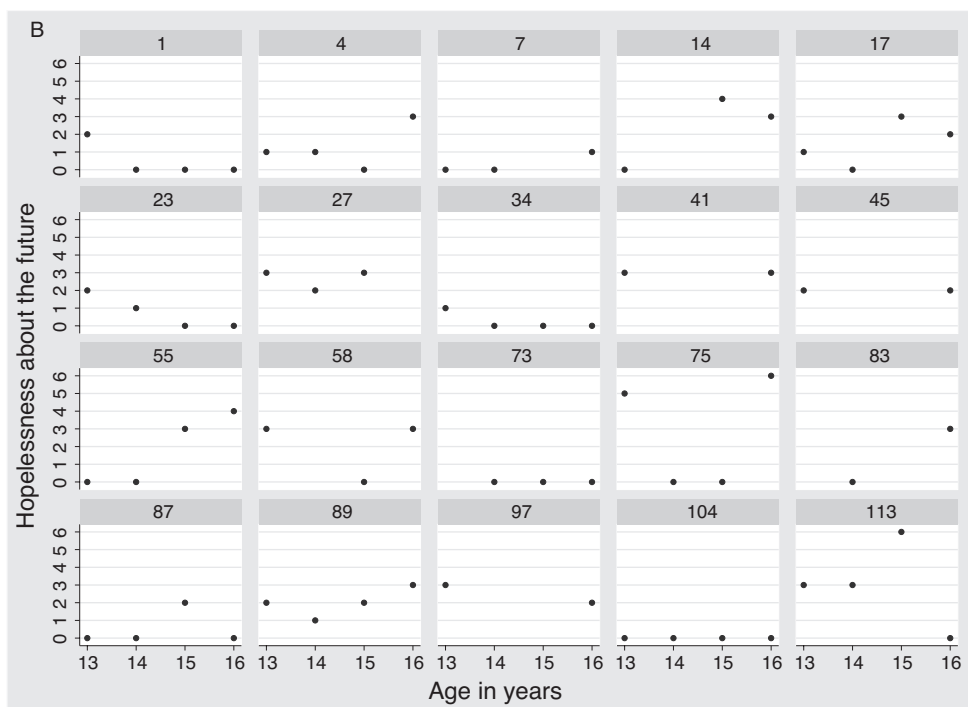
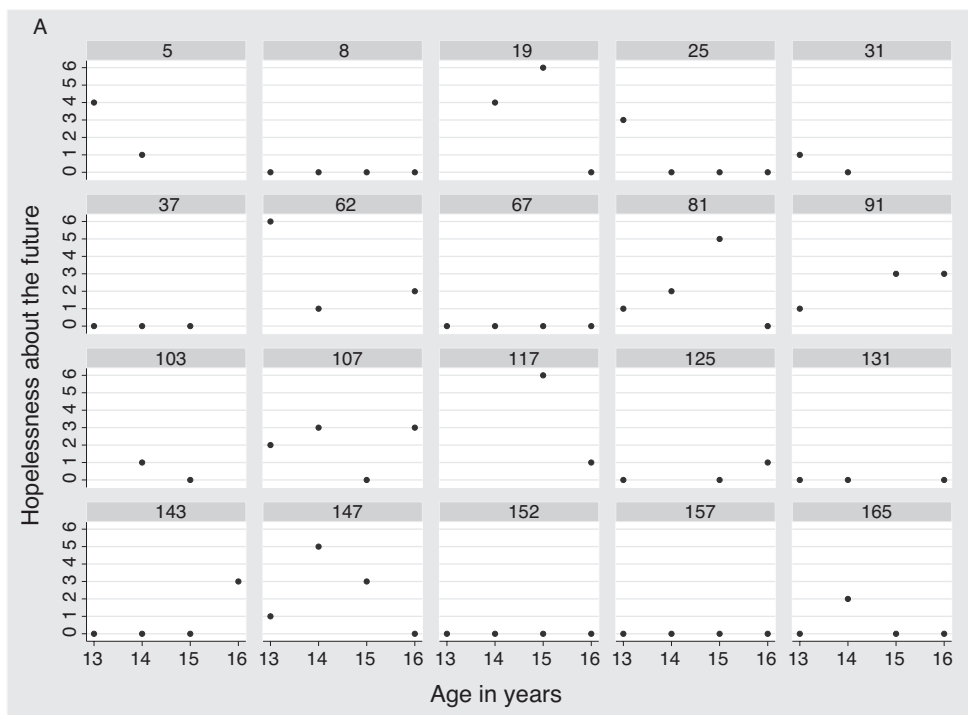


Figure 9.2 Scatterplots: individual hopelessness measurements over time. Time is indexed by age during middle adolescence. Panel A (top): Girls. Panel B (bottom): Boys.

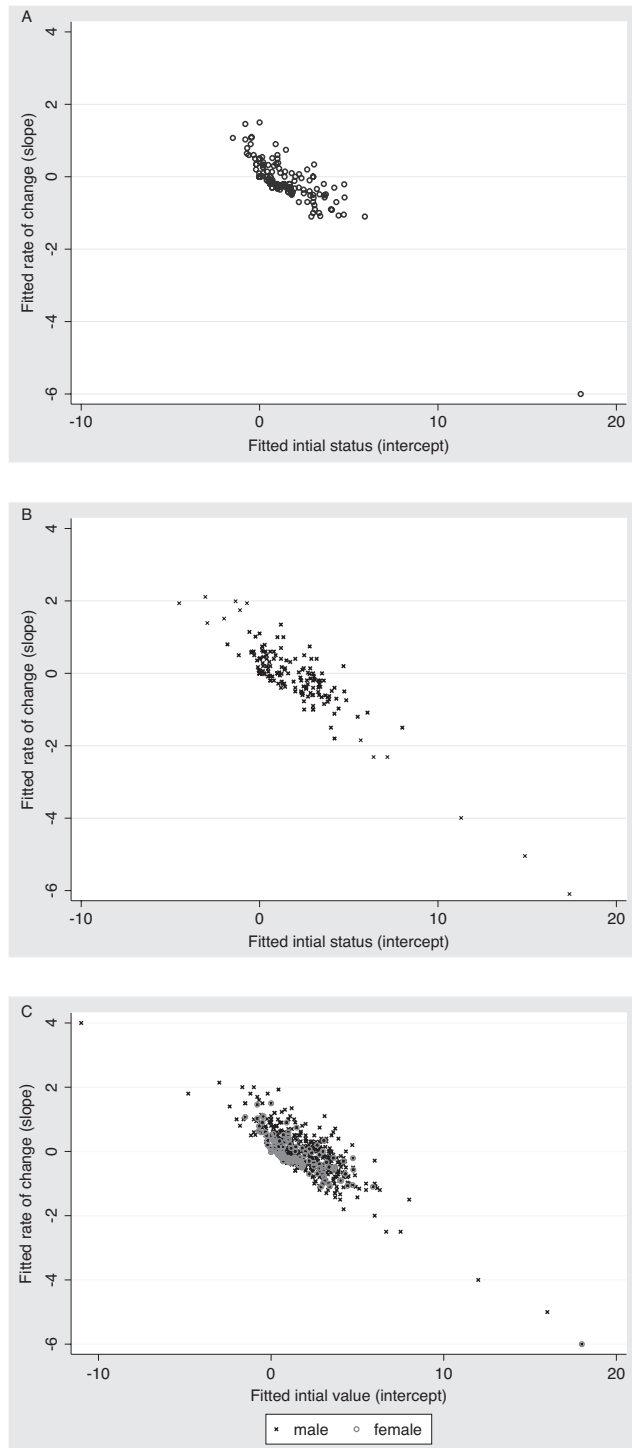


Figure 9.3 Scatterplots: ordinary least squares intercepts and slopes. A linear model for change was estimated. Panel A (top): girls; Panel B (middle): boys; Panel C (bottom): entire sample.

Table 9.1 Initial data analysis: OLS linear function for change

<i>Sample</i>	<i>Initial status (Intercept)</i>	<i>Rate of change (Slope)</i>
Total sample		
Mean	1.44	-.05
Standard deviation	1.91	.67
Correlation: intercept, slope		-.77*
Girls		
Mean	1.24	-.07
Standard deviation	2.00	.67
Correlation: intercept, slope		-.81*
Boys		
Mean	1.62	-.03
Standard deviation	1.80	.66
Correlation: intercept, slope		-.74*

Note. OLS = ordinary least squares.

* $p < .05$.

Based on these preliminary analyses, four theoretically plausible multiple group models were estimated simultaneously using gender as known groups:

- 1 Linear change with equal covariate prediction.
- 2 Linear change with class-specific covariate prediction.
- 3 Quadratic change with equal covariate prediction.
- 4 Quadratic change with class-specific covariate prediction.

The models included early adolescent covariate predictors (connection to mother, positive neighborhood connectedness, early fighting, and early violence with a weapon) and the distal outcome of violence with a weapon during later adolescence. Covariates were allowed to influence latent growth parameters (intercept, slope, and quadratic coefficients, as applicable), as well as class membership. As an example, Figure 9.4 shows *Mplus* syntax for the model for quadratic change with class-specific covariate effects for two trajectory classes each for boys and girls.

In our example, all participants with five or more time points (once during early adolescence, once during late adolescence, and at least three times during middle adolescence), regardless of start time or other missing measurement occasions, were included in the GGMM analysis. After selecting the sample from the larger study dataset, the sample was examined for missingness. Data were treated as MAR (following earlier decisions, see Bolland et al., 2007). The maximum likelihood estimator was used.

Table 9.2 summarizes statistical fit for the final four models that were considered. Theoretically, this analysis was based on the proposition that an adolescent's trajectory of hopelessness is related to differences in levels of connectedness to mother and neighborhood during early adolescence, as well as differences in violence involvement during early adolescence. Model 2 best met this proposition by allowing for covariate prediction to vary by latent class, however, in also

Model 3. Quadratic change, covariates equal across latent classes.	
TITLE:	4 class multiple group quadratic model with predictors and distal outcome; covariates equal across latent class.
DATA:	File is c:\data\dissertation.dat;
VARIABLE:	names are pid gender mother nhpos nhheg hope13 hope14 hope15 hope16 weapon eviol efight; Missing are all (-9999); Usevariable are hope13 hope14 hope15 hope16 mother nhpos weapon eviol efight; Categorical = weapon; Classes = gender (2) c (2); Knownclass = gender (gender = 0 gender = 1)
Analysis:	Type = mixture; Starts = 2000 20; Algorithm = integration EM;
Model:	%overall% i s q hope13@0 hope14@1 hope15@2 hope16@3; i s q on mother nhpos efight eviol; c on gender mother nhpos efight eviol;
Output:	Sampstat tech1 tech4 tech7;
Plot:	Type is plot3; Series = hope13 hope14 hope15 hope16 (*);

Figure 9.4 Mplus syntax: general growth mixture model. The syntax for the hope/hopelessness model specifies quadratic change and equal covariate effects across latent classes.

Table 9.2 Statistical fit for proposed multiple group models

Quality/index	Model 1 ^a	Model 2 ^b	Model 3	Model 4 ^c
Change model	Linear	Linear	Quadratic	Quadratic
Covariate effects	Equal	Class specific	Equal	Class specific
AIC	10972	10971	10844	NA
BIC	11114	11227	11050	NA
Sample Adj. BIC	11015	11049	10907	NA
Entropy	.871	.920	.908	
Number in class	C1 = 277	C1 = 295	C1 = 292	
	C2 = 74	C2 = 56	C2 = 59	
	C3 = 236	C3 = 113	C3 = 235	
	C4 = 134	C4 = 257	C4 = 135	

Note: Adj. = adjusted; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

^aIntercept and slope variance fixed at 0;

^bSlope variance fixed at 0;

^cEliminated latent variable covariance matrix (psi) not positive definite due to high correlations between the slope and quadratic coefficient. From Stoddard, Henly, Sieving, and Bolland (2011). Used with permission.

considering our statistical fit indices, Model 3 – in which the covariate prediction of connection to mother, positive neighborhood connection, early fighting and early violence with a weapon was equal for all classes – remained the model of choice. In addition, we proposed that the level and rate of change in hopelessness may vary during adolescence. A linear model of change suggests a constant rate of growth or decline in hopelessness during adolescence, whereas including a quadratic function suggests that hopelessness has a variable rate of change during middle adolescence. For example, an increasingly hopeless individual becomes increasingly hopeless faster over time. Based on statistical fit provided by the information criteria and entropy, Model 3 was most satisfactory. Model 3 (quadratic change with equal covariate prediction) was the best fitting model, as indicated by the lowest AIC, BIC, and sample adjusted BIC. In addition, Model 3 entropy was greater than 0.90. See Figure 9.5 for a path model depiction. Table 9.3 contains parameter estimates.

Two latent classes emerged for both boys and girls: A consistently low hopelessness class characterized by lower levels of hopelessness initially that continued to decrease over time, and an increasingly hopeless class characterized by higher levels of hopelessness initially that decreases slightly at age 14, then accelerated with time (Figure 9.6). Estimated odds ratios for violence with a weapon during later adolescence by latent hopelessness classes are listed in Table 9.4. The probability of violence with a weapon was higher in the increasingly hopeless classes compared to the low hopelessness classes. In addition, the probability of violence with a weapon during late adolescence was higher for boys than girls.

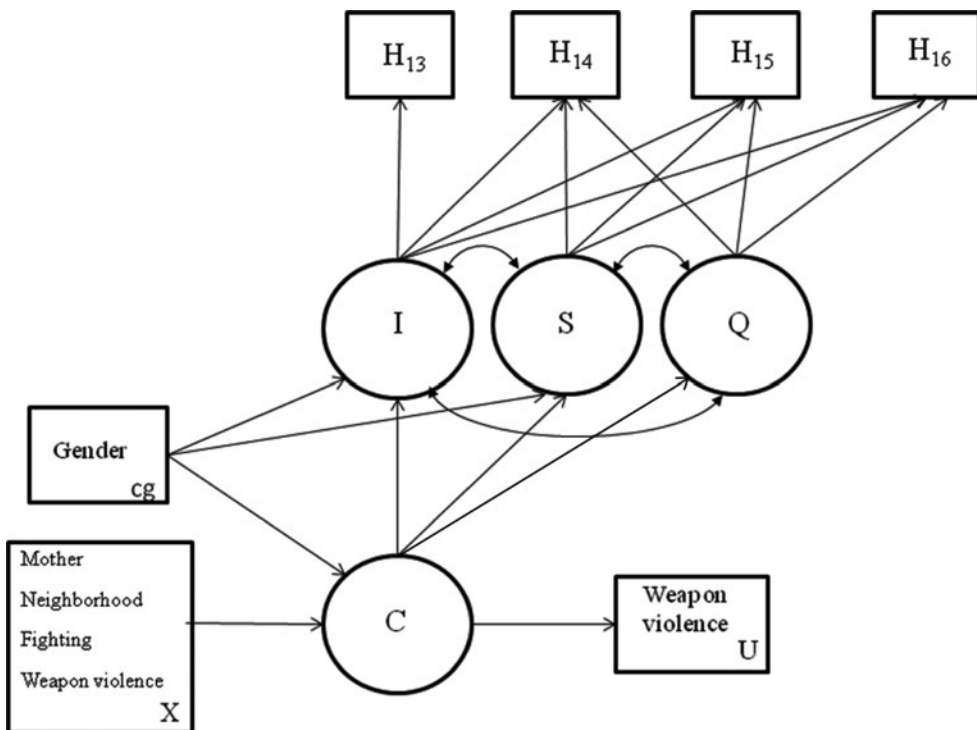


Figure 9.5 General growth mixture model: path model. This path model depicts Model 3: quadratic change, covariate effects equal across latent classes.

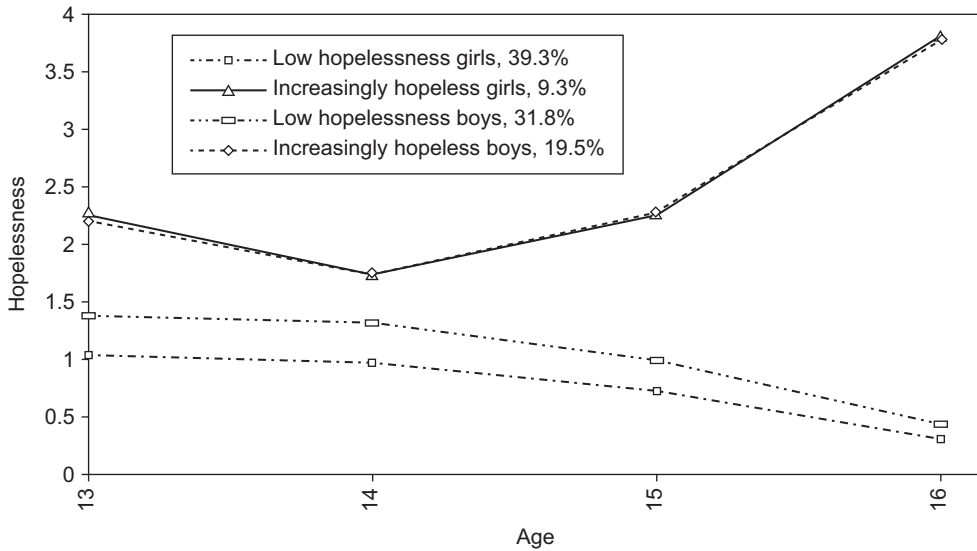


Figure 9.6 Estimated hopelessness trajectories. Final multiple group model of mean trajectories for hopelessness classes over time, adjusted for the effects of the covariates: low hopelessness girls (Class 1); increasingly hopeless girls (Class 2); low hopelessness boys (Class 3); and increasingly hopeless boys (Class 4). From Stoddard, Henly, Sieving, and Bolland (2011). Used with permission.

Table 9.4 Estimated odds ratios for violence with a weapon during later adolescence by latent hopelessness class

Comparison	Reference	OR	p
Low hopelessness girls	Increasingly hopeless girls	.59	.002
Low hopelessness girls	Low hopelessness boys	.44	.001
Low hopelessness girls	Increasingly hopeless boys	.40	.001
Increasingly hopeless girls	Low hopelessness boys	.74	.002
Increasingly hopeless girls	Increasingly hopeless boys	.68	.003
Low hopelessness boys	Increasingly hopeless boys	.92	.001

Note: OR = odds ratio (estimated).

Conclusion

This chapter provided an introduction to general growth mixture modeling (GGMM). It described and explained the methodology behind GGMM and used an example to demonstrate the modeling process step-by-step. It showed how GGMM can be used to assess similarities in interindividual differences in intraindividual change in health behavior or outcomes over time. The chapter provided an overview of model estimation, model evaluation and selection, and assessment of model fit. With careful attention to theory, design, and technical statistical issues, the GGMM can be useful in nursing science for understanding complex temporal processes involving latent classes.

References

- Acock, A. (2009, July). *Latent growth curve modeling using Mplus*. Summer Institute on Research Methodology. Unpublished document. Corvallis, OR: College of Health & Human Sciences, Oregon State University.
- Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, 8, 338–363.
- Bolland, J. M., Bryant, C. M., Lian, B. E., McCallum, D. M., Vazsonyi, A. T., & Barth, J. M. (2007). Development and risk behavior among African American, Caucasian, and mixed-race adolescents living in high poverty inner-city neighborhoods. *American Journal of Community Psychology*, 40, 230–249.
- Bolland, J. M., Lian, B. E., & Formichella, C. M. (2005). The origins of hopelessness among inner-city African-American adolescents. *American Journal of Community Psychology*, 36, 293–305.
- Bolland, J. M., McCallum, D. M., Lian, B., Bailey, C. J., & Rowan, P. (2001). Hopelessness and violence among inner-city youths. *Maternal and Child Health Journal*, 5, 237–244.
- Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge: Cambridge University Press.
- Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, 57, 505–528.
- Cudeck, R., & Henly, S. J. (2003). A realistic perspective on pattern representation in growth data: Comment on Bauer and Curran (2003). *Psychological Methods*, 8, 378–383.
- Farrington, D. P., Langan, P. A., & Tonry, M. (2004). *Cross-national studies in crime and justice*. Washington, DC: Bureau of Justice Statistics.
- Henly, S. J., Vermeersch, P. E. H., & Duckett, L. J. (1998). Model selection for covariance structures analysis in nursing research. *Western Journal of Nursing*, 20, 344–355.
- Henly, S. J., Wyman, J. F., & Findorff, M. J. (2011). Health and illness over time: The trajectory perspective in nursing science. *Nursing Research*, 60(Suppl. 3), S5–S14.
- Li, F., Duncan, T. E., Duncan, S. C., & Acock, A. (2001). Latent growth modeling of longitudinal data: A finite growth mixture modeling approach. *Structural Equation Modeling*, 8, 493–530.
- Loeber, R., & Stouthamer-Loeber, M. (1998). Development of juvenile aggression and violence: Some common misconceptions and controversies. *American Psychologist*, 53, 242–259.
- Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class-latent growth modeling. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 291–322). Washington DC: American Psychological Association.
- Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 345–368). Thousand Oaks, CA: Sage.
- Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S.-T., Yang, C.-C., . . . Liao, J. (2002). General growth mixture modeling for randomized preventative interventions. *Biostatistics*, 3, 459–475.
- Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55, 463–469.
- Muthén, B. O. (1998–2004). *Mplus technical appendices*. Los Angeles, CA: Author.
- Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Author.
- Nagin, D. S., & Odgers, C. L. (2010). Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology*, 6, 109–138.
- Nies, M. A., & Sun, Y. (2008). Responders and nonresponders to a walking intervention for sedentary women. *Journal of Nursing Scholarship*, 40, 226–234.
- Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, 52, 501–525.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.
- Stein, N. R., Dickstein, B. D., Schuster, J., Litz, B. T., & Resick, P. A. (2012). Trajectories of response to treatment for posttraumatic stress disorder. *Behavior Therapy*, 43, 790–800.
- Stoddard, S. A., Henly, S. J., Sieving, R. E., & Bolland, J. (2011). Social connections, trajectories of hopelessness, and serious violence in impoverished urban youth. *Journal of Youth and Adolescence*, 40, 278–295.
- Titterton, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester: Wiley.

- Tremblay, R. E., Nagin, D. S., Séguin, J. R., Zoccolillo, M., Zelazo, P. D., Boivin, M., . . . Japel, C. (2004). Physical aggression during early childhood: Trajectories and predictors. *Pediatrics*, *114*, e43–e50.
- Tseng, C.-H., Elashoff, R., Li, N., & Li, G. (2012). Longitudinal data analysis with non-ignorable missing data. *Statistical Methods in Medical Research*. Advance online publication. Retrieved from www.ncbi.nlm.nih.gov/pmc/articles/PMC3883866/
- Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, *91*, 217–221.
- Wills, A. K., Lawlor, D. A., Muniz-Terrera, G., Matthews, F., Cooper, R., Ghosh, A. K., & Hardy, R. (2012). Population heterogeneity in trajectories of midlife blood pressure. *Epidemiology*, *23*, 203–211.

10

MULTILEVEL MODELS

Ulrike Grittner and Nils Lahmann

This chapter discusses how and why multilevel modeling is a flexible and powerful tool for analyzing data with a hierarchical structure. Such data are often found in nursing science and public health research (e.g., when analyzing patients in clinics or hospitals). Using example datasets, we explain step-by-step how to conduct linear multilevel modeling. Additionally we provide syntax commands for several software packages and demonstrate how to interpret the results of multilevel analyses.

The aim of multilevel modeling is to integrate the regression equation at a lower level of data grouping (usually individuals) with that at higher levels (such as hospitals) into one regression equation and to incorporate covariates at appropriate levels. By using multilevel models it is possible to adjust for similarity of the lower-level units belonging to the same group of a higher level, and to make overall inferences about relationships between lower level as well as higher-level characteristics and the outcome of interest.

Key ideas

In social research, studies that investigate the relationship between individuals and the society they live in are very common. Individuals interact within the social context of their group; i.e., they are influenced by group and contextual characteristics and, in turn, group characteristics are influenced by the individuals who constitute the group. As a result of group clustering, data from individuals of a group are often more similar to each other (i.e., correlated) in contrast to data from individuals from different groups.

Such clustered or hierarchical data can be analyzed at different hierarchical levels, while variables at each level may be incorporated. This leads to research that identifies the variance in the outcome that is related to the different levels of data and at the appropriate levels also evaluates the association of individual or group characteristics with the outcome. This kind of analysis is referred to as “multilevel analysis” (Snijders & Bosker, 1999). The method has been described under various names such as: hierarchical models (Raudenbush & Bryk, 2002), random effects models (Dunson, 2008), and mixed models (Verbeke & Molenberghs, 2000; Zuur, Ieno, Walker, Saveliev, & Smith, 2009).

Multilevel modeling has become a useful analytic tool not only in the social sciences and clinical sciences, but also within epidemiological, public health, and health promotion research (e.g., Diez-Roux, 2000; Diez-Roux & Aiello, 2005; Kothari & Birch, 2004). This technique has expanded the range and depth of analyses such that researchers may now take into account more than one level of data. This is, for example, relevant for research in health promotion that is often based on a socio-ecological theoretical framework (Kothari & Birch, 2004). Influencing the health of individuals may occur not only directly but also via the families, neighborhoods and communities in which they live, the places at which they work, and the schools that they attend. Recent research has employed multilevel modeling to examine a variety of subjects including the effect of the community on the rate of unplanned pregnancies (Koren & Mawn 2010), the effects of the social and built environment on stress and health (Matthews & Yang, 2010), and the effects of fast-food outlet density and car ownership on body mass (Inagami, Cohen, Brown, & Asch, 2009). The technique has also found wide use in intervention research including such recent studies as one of blood pressure reduction among African-American men visiting barber shops (Victor et al., 2009), an evaluation of a web-based anti-smoking campaign in Toronto secondary schools (Norman, Maley, Li, & Skinner, 2008), and a trial to reduce stress among young restaurant workers (Petree, Broome, & Bennett, 2012).

It is thus clear that multilevel approaches have become well accepted in the health research field. They add an extra “dimension” of analysis to produce results that more fully take into account additional avenues of influence upon human health and behavior. With respect to nursing science and clinical epidemiology, most studies of treatment effects on health aim at controlling individual factors (such as age, sex, educational achievement, income) in the analysis. This requires multilevel modeling due to the fact that data nested at different levels of sampling are combined into one analysis. Some examples of hierarchical data include patients who are nested in different clinical specialties (intensive treatment unit, surgery, etc.) or in clinics or hospitals, residents who live in long-term care facilities, nurses who are nested in wards, teeth that are nested in Question: the mouth of a patient, and a prominent example: repeated measurements are nested in subjects. Here we name the lower-level units (individuals, patients, residents, teeth, nurses, repeated measurements,) as level 1 units and the higher-level units (specialties, hospitals, wards, mouth of one patient, subjects) as level 2 units.

In all these cases there might be a dependence or correlation between the level 1 units that cannot be ignored with regard to the analyzed outcome. It makes a difference if analyzing, for example, data on 300 patients from 300 different wards (no hierarchical data structure) in contrast to an analysis of data on 300 patients from 15 different wards (20 patients per ward). Patients on one ward tend to be more similar to each other than to patients in different wards with regard to an outcome variable such as risk for developing pressure ulcers. To ignore this dependency may yield misleading results.

We assume that the reader is familiar with linear regression. For background reading about linear regression, we recommend for example Montgomery and Peck (1992). We will not explain here the more technical details or the mathematical background of the models used. Using a small data example, we will discuss how traditional techniques fail to model the data in an appropriate way and will then show how a linear multilevel model is applied. We will then analyze the whole data set. We will provide software commands in different software languages. Within the multilevel framework it is also possible to fit a logistic regression, Poisson regression, or other regression models that we will not discuss here because the idea of multilevel modeling is the same for all these regression models. Specific applications of these extensions to the basic model are discussed in Snijders and Bosker (1999).

The data example

As an illustrative example, we use data from the German hospital prevalence study about pressure ulcers (Kottner, Wilborn, Dassen, & Lahmann, 2009; Lahmann, Halfens, & Dassen, 2005; see Table 10.1). The data set consists of information from 11,057 patients from 22 hospitals in the years 2008 to 2011. The Braden scale (Bergstrom, Braden, Laguzza, & Holman, 1987) was filled in for each participant to assess the individual pressure ulcer risk. Based on a conceptual scheme for PU development this instrument consists of six items. The items “sensory perception,” “moisture,” “activity,” “mobility,” and “nutrition” are rated from 1 (*most impaired*) to 4 (*least impaired*). The item “friction and shear” is rated from 1 (*problem*) to a maximum of 3 (*no apparent problem*). Item scores are summed up. Thus, sum scores from 6 (*maximum pressure ulcer risk*) to 23 (*no pressure ulcer risk*) were generated. In this study, the cut-off point for pressure ulcer risk was set at a maximum of 20 for all patients to exclude those patients with no risk of pressure ulcer at all resulting in a data set containing information of 3,956 patients.

In our example we focus on the relationship between individual (level 1) and hospital characteristics (level 2) on one hand and the Braden score of the patients (the outcome variable) on the other hand. The rationale for the multilevel approach to the study was that the Braden scale assessment was done by the individual hospital staff themselves. Thus the training for the Braden risk assessment was done by local coordinators of each participating hospital. The quality of this training may have differed between the participating institutions. Further some hospitals had used the Braden score already in their clinical care protocols while others only used it for the study. These differences may lead to different Braden sum score results in patients who have the same amount of risk. Multilevel modeling takes these differences into account and quantifies hospital related differences in Braden scores after adjusting for individual characteristics of the patients.

Table 10.1 Descriptive data for 10 selected hospitals

Hospital	N	Age		Braden score	
		M	(SD)	Mdn	(IQR)
1	127	75.8	(11.4)	17	(14–19)
2	141	72.6	(14.8)	17	(15–19)
3	393	73.2	(14.0)	17	(14–19)
4	255	73.4	(13.3)	18	(15–19)
5	277	70.1	(15.0)	18	(14–19)
6	269	70.1	(15.7)	17	(13–19)
7	110	65.5	(14.8)	18	(17–20)
8	108	57.8	(15.5)	18	(16–19)
9	70	67.3	(18.6)	15	(12–18)
10	225	74.6	(13.1)	18	(15–19)

Note. N = number of patients interviewed at each hospital.

M = mean. SD = standard deviation.

Mdn = median. IQR = interquartile range.

The Braden score ranged between 6 and 20 in the sample with a mean of 16.5 ($SD = 3.3$) and a median of 17. The distribution of Braden score values in the sample is left skewed (skewness: -0.93). To reduce skewness we transformed the values first by dividing them by 10 and second by calculating the fourth power resulting in a distribution with a skewness of -0.10 . The transformed values are used in the examples.

For the following examples (except the last), we use first a subsample of 10 selected hospitals containing information of Braden score and age from 1,975 patients from this larger data set. We wish to examine the relationship between age and Braden score.

Traditional techniques and their limitations

Regression at patient level – ignoring nested data structure and dependence

We could analyze these data using traditional statistical techniques such as ordinary least squares linear regression (Kreft & De Leeuw, 1998). Linear regression models have some assumptions. These are: normally distributed residuals after applying the model, homogeneity in the variance of the outcome over all covariates, fixed covariate matrix, independence among the analyzed units (individuals), and correct model specification.

With an ordinary linear regression we would ignore the fact that patients are nested in hospitals. Yet, patients in one hospital might be more similar with regard to their Braden scores than patients in another hospital (for example depending on the specification of the hospital) and the relationship between age and Braden score might differ between hospitals. In fact the range between the 10 hospitals with regard to the median Braden scores of the patients is between 15 and 18 (Table 10.1) and with a Kruskal-Wallis test we would conclude that there are significant differences with regard to the Braden scores between the 10 hospitals ($p < .001$). Further, in eight of the hospitals (1, 2, 3, 4, 5, 7, 8, 10), higher age of the patients is related to lower Braden scores while in two hospitals (6, 9) the opposite is true. Therefore we cannot assume that the nested structure of the data is unimportant. Taking this into consideration would be a better use of all the information that the dataset provides and would offer more explanatory power regarding the Braden scores.

Hospital level regression with aggregated data – ignoring within-hospital variance

If we are more interested in the differences between *hospitals* with regard to Braden scores than in the differences between *individual patients*, we could simply use aggregated data from the 10 hospitals. Here we would use the mean of the transformed Braden scores as outcome and the mean age (centered at 65 years) of patients as covariate.

With an aggregated regression, however, we ignore variability within the hospitals (i.e., among the patients). This regression is based only on 10 units of observation (hospitals). The results must be interpreted cautiously because we cannot say anything about the individuals (i.e., how individual age would be related to individual Braden score). Any conclusions about individuals based on hospital-level analysis risks the inferential problem that is known as the ecological fallacy (Freedman, 2004).

Analysis of covariance at patient level – accounting for different hospitals

With an analysis of covariance (ANCOVA) we can test the relationship between age of patients and Braden score while accounting for differences between hospitals with regard to average Braden score. If we additionally include an interaction term for hospitals and age we can test if the relationship between age and Braden score differs between hospitals. If the interaction term is significant, it is more appropriate to calculate 10 individual regressions with an ordinary least square estimation instead of ANCOVA analysis because otherwise the *F*-tests are biased (see, for example, Bijleveld & van der Kamp, 1998; Kreft & de Leeuw, 1998).

The advantage of ANCOVA or separate regressions for every hospital is that we model more adequately the differences between hospitals with regard to mean Braden scores as well as the relationship between patient age and Braden scores. This is in contrast to a total or aggregated regression where we have only one overall (over all patients from all hospitals) slope coefficient for the relationship between age and Braden score.

However, there are some drawbacks to ANCOVA or separate regressions including that ANCOVA indicates that differences exist between hospitals but we cannot test or further explain why differences between hospitals exist. It is not possible to incorporate hospital-level characteristics as covariates on the hospital level in the model. Another disadvantage is the large number of parameter estimates. Even if there would not be significant interaction we would still obtain 11 regression parameters (10 intercepts for the hospitals and a slope parameter for age) and this number increases with every additional level-2 unit. That means that ANCOVA or separate regressions are only reasonable if the number of level-2 units is small (<20).

Multilevel modeling

The main idea of multilevel modeling is to integrate the regression at level 1 (e.g., individuals) with level 2 (e.g., hospitals) or higher levels into one regression equation and also to incorporate covariates at appropriate levels. In addition, it is possible to adjust for correlation among level-1 units; i.e., similarity of the level-1 units that belong to the same level-2 unit. Alternatively, we could simply conduct extra regressions for each cluster. But if we want to make overall inferences about relationships between lower level as well as higher-level characteristics and the outcome, we can combine these regressions into one regression model with multilevel techniques.

The assumptions underlying the multilevel linear regression model are similar to those for ordinary multiple regression analysis: linear relationships, homoscedasticity, and normal distribution of the residuals (Maas & Hox, 2004). But now dependence of lower-level units is also accounted for.

Building the multilevel model

File structure

For a multilevel analysis most software packages require that the data be arranged in the following way (Figure 10.1): one line (row) for every level-1 unit (e.g., patients: ‘patient_id’), a variable for identification of higher-level units (e.g., hospitals: here ‘hospital’ with a hospital id number), covariates that contain information about level 1 or higher-level characteristics as variables with entries in every line (here, Braden score and age for each patient). Once the file is set up, multilevel linear regression modeling can proceed step-by-step as demonstrated in the following sections.

hospital	patient_id	Braden	age
1	1	18	78
1	2	9	75
1	3	10	89
1	4	10	85
.	.	.	.
.	.	.	.
.	.	.	.
2	5	17	77
2	6	16	83
2	7	15	87
2	8	14	66
.	.	.	.
.	.	.	.
.	.	.	.
3	9	10	68
3	10	20	67
3	11	19	75
.	.	.	.
.	.	.	.
.	.	.	.

Figure 10.1 Example of multilevel data structure.

Step 1: Random intercept model without covariates

In the first step we analyze a model with no explanatory variables. This model is called the intercept-only model. The equation is

$$y_{ij} = \beta_0 + u_{0j} + \varepsilon_{ij}, \quad (10.1)$$

where

y_{ij} is the transformed Braden score value of patient i in hospital j ,

β_0 is the intercept,

u_{0j} is the residual or random effect for hospital j (mean = 0, variance = σ_{u0}^2), and

ε_{ij} is the error term for patient i at hospital j (mean = 0, variance = σ_{ε}^2).

This model does not explain factors that are related to the Braden score of patients. It only decomposes the variance of Braden scores into a component that is related to hospital differences σ_{u0}^2 and a component that is related to differences between patients σ_{ε}^2 .

The SAS (SAS System for Windows, 2008) syntax to run this model is:

```
proc mixed data=brad covtest ;
    class hospital ;
    model Braden= /Solution cl ;
    random intercept / subject=hospital;
run;
```

where *brad* is the name of our data set, *hospital* is a variable that contains one specific numeric value for every hospital, and *Braden* is the variable that contains the values for the Braden scores of the patients.

In the Appendix we list the corresponding commands in STATA (StataCorp., 2007), R (Pinheiro, Bates, DebRoy, Sarkar, & R Development Core Team, 2011; R Development Core Team, 2011), and SPSS (SPSS for Windows, 2011). We will not explain in detail here what the commands stand for. For further details see Venables and Ripley (2002), Verbeke and Molenberghs (2000), or Zuur et al. (2009). Note: for the random effects in STATA and R, the standard deviation is given while in SAS and SPSS the variance (the square of the standard deviation) is given in the output.

With Equation 10.1 for the intercept-only model, we are interested in the intercept β_0 and the two variance estimates: σ_{u0}^2 for the variance between hospitals and σ_ϵ^2 for the variance between patients. From the output we obtain the values depicted in Table 10.2.

“Fixed effects” are the overall effects across all hospitals and can be interpreted in the same way as in non-hierarchical linear regression. Here the fixed effect is the parameter for the intercept (β_0). But we know already that the hospitals differ with regard to the Braden scores of the patients. Especially when we analyze a sample with many hospitals, we are not interested in the particular intercepts and effects from each hospital, but in the *distribution* of intercepts across hospitals, because then we see if patients in different hospitals are more or less similar with regard to the outcome. Thus we examine the variance estimate for the hospital intercepts (σ_{u0}^2). It is significant here, meaning that there is significant variation among the 10 hospitals with regard to the mean (transformed) Braden score of their patients.

The value for σ_ϵ^2 is a measure for the residual variance on the patient level and is significant, meaning that there are significant differences with regard to Braden scores among the individual patients after taking into account differences in mean (transformed) Braden scores between hospitals.

We can now calculate the intraclass correlation (ICC), which is a measure of the average degree of dependence of level-1 units (patients) within a hospital. Here we obtain an estimate for the average similarity or dependence of patients within a hospital with regard to Braden scores. The equation for calculation of the ICC is

$$\bar{\rho} = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_\epsilon^2} \quad (10.2)$$

It is the proportion of variance that is accounted for by the hospitals.

Table 10.2 Multilevel linear model for transformed braden scores: random intercept model without covariates

<i>Effects</i>	<i>Parameter</i>	<i>Estimate</i>	(SE)
Fixed	Intercept β_0	8.95	(0.33)
Random	Variance between hospitals σ_{u0}^2	0.95	(0.57)
	Variance between patients σ_ϵ^2	25.30	(0.81)

Note. Results are based on 1,975 patients in 10 hospitals. Significant effects are shown in bold.

In our example the estimated ICC is $\bar{\rho} = \frac{0.95}{0.95 + 25.30} = 0.04$. This means that 4% of the variance in patient Braden scores is explained by hospital differences. We conclude that differences of Braden scores between patients are present but are only explained to a small degree by hospital differences.

Step 2: Integrating covariates at the patient level as fixed effects

Now we want to extend our model to see how age is related to Braden scores.

Our model can be described by the following equation:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + \varepsilon_{ij}, \quad (10.3)$$

where

y_{ij} is the transformed Braden score value of patient i in hospital j ,

β_0 is the fixed effect intercept,

β_1 is the fixed effect slope for age,

x_{1ij} is the age (centered at 65, divided by 10) of patient i in hospital j ,

u_{0j} is the residual or random effect for the intercept for hospital j (mean = 0, variance = σ_{u0}^2), and

ε_{ij} is the error term for patient i at hospital j (mean = 0, variance = σ_ε^2).

To estimate parameters of this model, we modify the previous syntax by adding to the command line “age.” In SAS:

```
proc mixed data=brad covtest ;
    class hospital ;
    model Braden = age /Solution cl ;
    random intercept / subject=hospital;
run;
```

As the results of the analysis show (Table 10.3), the slope coefficient for age (β_1) is negative and significant, meaning that those patients of higher age have lower Braden scores and are at greater PU risk.

However, the regression-related graph (Figure 10.2, Panel B) looks very different compared to Figure 10.2, Panel A. The separate regressions (Figure 10.2, Panel A) describe our data more

Table 10.3 Multilevel linear model for transformed braden scores: random intercept model with individual level covariate

Effects	Parameter	Estimate	(SE)
Fixed	Intercept β_0	9.18	(0.32)
	Slope for age (centred in decades) β_1	-0.42	(0.08)
Random	Variance between hospitals intercepts σ_{u0}^2	0.82	(0.51)
	Variance between patients σ_ε^2	24.94	(0.80)

Note: Results are based on 1,975 patients in 10 hospitals. Significant effects are in bold.

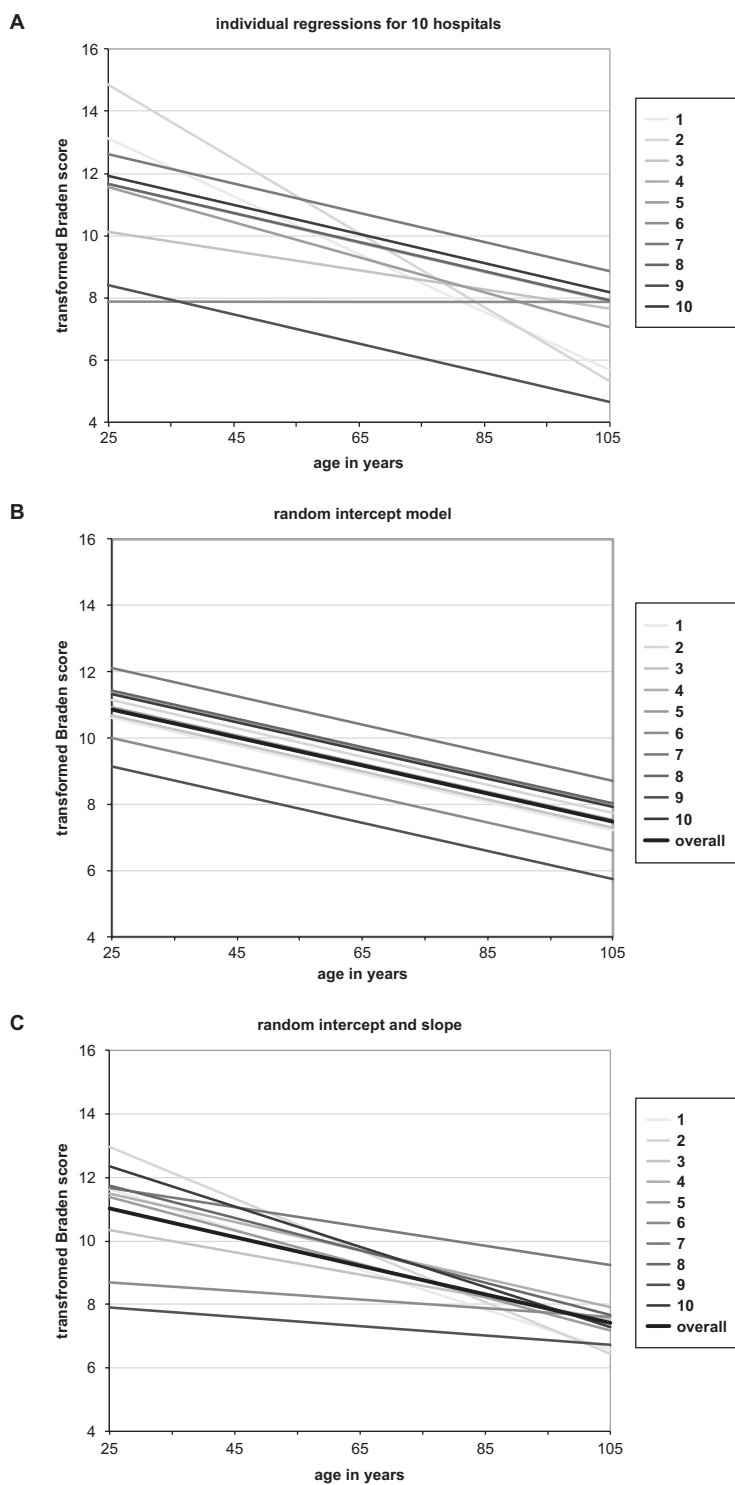


Figure 10.2 Regression of transformed Braden scores on age using three different models.

exactly because in the random intercept model we have assumed that the relationship between age and Braden scores (the slope) is the same in each hospital, which is not exactly the case. We now want to test if the differences between hospitals with regard to the relationship between age and Braden scores (the slopes) are significant.

It should be noted that other covariates on the individual level could also be integrated in the model. When analyzing data sets with more units on the higher level, additional covariates on that level could also be tested, for example, different hospital characteristics.

Step 3: Adding random slopes

We will now add a random slope term for age to the model. That means we allow not only for the intercepts to vary between hospitals but also the slopes. The equation is now extended by $u_{1j} x_{ij}$:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j} x_{ij} + \varepsilon_{ij} \quad (10.4)$$

Here $\beta_1 + u_{1j}$ is the slope for age in hospital j , where β_1 is the fixed part and u_{1j} is the random part with mean = 0 and variance σ_{u1}^2 . To estimate in SAS, we specified:

```
proc mixed data=brad covtest ;
    class hospital ;
    model Braden = age /Solution cl ;
    random intercept age / subject=hospital;
run;
```

As indicated by the fixed effect for the slope (β_1), age is significantly related to Braden scores (Table 10.4). Higher age is associated with lower Braden scores (greater pressure ulcer risk). The variance coefficient for the intercept (σ_{u0}^2) now shows that there are significant differences between hospitals with regard to the intercepts, meaning the mean Braden scores for the patients in the different hospitals differ significantly. But the variance coefficient for the slopes (σ_{u1}^2) is not significant meaning that the relation between age and Braden scores in the 10 hospitals is not significantly different. Figure 10.2, Panel C shows the regression-based lines for the 10 hospitals. When comparing Figure 10.2, Panel C and Figure 10.2, Panel A it is obvious that the real slopes are much steeper and that the intercepts and slopes from the separate regressions differ more than those from the multilevel model.

The reason for this is due to the weighting of the regression coefficients of the 10 hospitals, which is called “shrinkage.” The “shrinkage” method, which is present in every multilevel regression, is a

Table 10.4 Multilevel linear model for transformed braden scores: random intercepts and slopes model with individual level covariate (10 hospitals)

Effects	Parameter	Estimate	(SE)
Fixed	Intercept β_0	9.22	(0.34)
	Slope for age (centred, in decades) β_1	-0.45	(0.12)
Random	Variance between hospital intercepts σ_{u0}^2	0.98	(0.59)
	Variance between hospital slopes σ_{u1}^2	0.09	(0.08)
	Variance between patients σ_ε^2	24.78	(0.79)

Note. Based on 1,975 patients in 10 hospitals. Significant effects are shown in bold.

method that “borrows strength” for the estimates of a particular hospital from the overall estimates (Kreft & De Leeuw, 1998). This prevents our variance estimate on the second level to be affected by outliers and produces more precise estimates (i.e., the standard errors are smaller). Thus, individual estimates of single hospitals are “shrunk” in the direction toward the overall estimate. The “shrinkage estimates” (so-called Empirical Bayes estimates) are weighted averages of the specific ordinary least squares estimate in each hospital and the overall regression coefficient (Hox, 2002). The extent of shrinkage depends on the group size and the difference between hospital-specific estimates and overall estimates. If the group size is relatively small and the distance between hospital-based estimates to the overall estimate is relatively high, the shrinkage will be larger than for larger group sizes and lower distance of hospital-specific estimates to overall estimates. The underlying assumption is that the hospital-specific estimates constitute a random sample from a normal distribution of such estimates. If particular group values are based on only a small number of patients, these could lead to extreme values, and to poor representation of the distribution of such estimates.

Calculations with the whole data set

So far we have worked with a small data subsample. We now calculate a series of models for the whole example data set with 22 hospitals and 3,956 patients: random intercept only, random intercept + covariate, random intercept + random slope; and random intercept + more covariates. The results from the different models (all the steps discussed so far) are shown in Table 10.5.

In the final model (Table 10.5, model 3, rightmost column), we have included additional covariates for urinary incontinence (yes/no), BMI (in kg/m²), and hospital departments (Internal medicine, Surgery, Intensive care, Neurology, Urology/Gynaecology, Psychiatry, Geriatrics, other). We also tested if there were significant differences between men and women with regard to Braden scores. But after adjusting for other covariates, there were no significant gender differences and therefore we did not include the covariate “sex” in the final model.

Variable selection methods can be used to find a final model. The Likelihood Ratio Test (LRT) compares the model fit and improvement of the model when adding covariates (Lehmann, 2005). In the example the test statistic shows that the final model is significantly better than the model with only one covariate (Table 10.5).

In the third column we tested a random slope for age (σ_{u1}^2), but because in the majority of hospitals older patients have lower Braden scores than younger patients, the slopes were similar and there was no significant variation between the slopes of the hospitals. Therefore we continued with the random intercept model.

The equation for the final model in the last column is:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_5 x_{5ij} + \beta_6 x_{6ij} + \beta_7 x_{7ij} + \beta_8 x_{8ij} + \beta_9 x_{9ij} + \beta_{10} x_{10ij} + u_{0j} + \varepsilon_{ij} , \quad (10.5)$$

where

y_{ij} is the (transformed) Braden score value of patient i in hospital j ,

β_0 is the fixed effect intercept,

β_1 is the fixed effect for age (centered at 65, in decades),

β_2 is the fixed effect for urinary incontinence (1: yes, 0: no),

β_3 is the fixed effect for BMI,

β_5 is the fixed effect for the Surgery department (1: surgery, 0: else),

β_6 is the fixed effect for the Intensive care department (1: intensive care, 0: else),

β_7 is the fixed effect for the Neurology department (1: neurology, 0: else),

β_8 is the fixed effect for the Urology/Gynecology (UG) department (1: UG, 0: else),

Table 10.5 Multilevel linear models for transformed braden scores: random intercepts and slopes models with individual-level covariates (22 hospitals)

Effects	Parameter	Model					
		0 Intercept Only (N = 3,956)		1 Intercept + Covariate (N = 3,956)		2 Intercept + Covariate + Random Slope (N = 3,956)	
		Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
Fixed	Intercept β_0	9.10	(0.17)	9.40	(0.16)	9.40	(0.17)
	Age (centred, in decades) β_1	—		-0.46	(0.06)	-0.46	(0.06)
	Urinary incontinence β_2					-4.26	(0.16)
	BMI β_3					0.05	(0.01)
	Sex β_4					—	
	Department (ref: Internal medicine)						
	Surgery β_5					0.07	(0.19)
Random	Intensive care β_6					-3.61	(0.27)
	Neurology β_7					-0.26	(0.26)
	Urology/Gynaecology β_8					1.26	(0.35)
	Psychiatry β_9					2.09	(0.36)
	Geriatrics β_{10}					1.30	(0.40)
	Other β_{11}					-0.30	(0.33)
	Variance between hospital intercepts $\sigma_{\mu_0}^2$	0.46	(0.22)	0.36	(0.19)	0.37	(0.20)
	Variance between hospital slopes for age $\sigma_{\alpha_1}^2$	—		—		0.01	(0.03)
	Variance between patients σ_e^2	25.36	(0.59)	24.93	(0.58)	24.90	(0.58)
	Intraclass Correlation (ICC)	0.02					
	Likelihood-Ratio-Test (Model 1 vs Model 3)					966.7^a	

Note. Models estimated using data from 22 hospitals. N = number of patients. Significant effects are shown in bold. ^a $df = 9, p < .001$.

β_9 is the fixed effect for the Psychiatry department (1: *Psychiatry*, 0: *else*),
 β_{10} is the fixed effect for the Geriatrics department (1: *Geriatrics*, 0: *else*),
 β_{11} is the fixed effect for the other departments (1: *Other*, 0: *else*),
 u_{0j} is the random effect for the intercept for hospital j (mean = 0, variance = σ_{u0}^2), and
 ε_{ij} is the error term for patient i at hospital j (mean = 0, variance = σ_ε^2).

Interpretation of the multilevel model

In addition to the age of the patients, urinary incontinence was negatively related to the Braden scores of patients (as indicated by significant coefficients β_1 and β_2). Older patients and those who were incontinent had lower scores than younger patients or those who were not incontinent. BMI was positively associated with Braden scores (as indicated by the coefficient β_3) meaning that patients with lower BMI were at higher risk for pressure ulcer than patients with normal weight, a phenomena that is already known (Drake et al., 2010). Patients from Intensive care departments had significant lower Braden scores than patients from Internal medicine departments while patients from Urology/Gynecology departments, Psychiatric departments, and Geriatrics departments had higher scores after adjustment for age and urine incontinence.

After adjusting for more individual characteristics, the variance between intercepts of hospitals and the residual variance of patients in the model were reduced. There were no significant random slope variables, indicating that the relationship between a covariate and the outcome (Braden score) did not differ significantly between hospitals (e.g., positive association in some hospitals, negative association in other hospitals). Further we could not find significant hospital-level variables that were related to the Braden scores of the patients. Therefore we might conclude that the Braden score of patients is related to individual characteristics of patients, not to the hospital characteristics that were analyzed here, and that the relationship between the individual characteristics and the Braden scores is similar in different hospitals. That means the influence of different qualities in the training of the Braden scale can be considered of minor relevance. However, as indicated by the significant variance coefficient for the random intercept, there are significant differences between hospitals with regard to the mean Braden scores of the patients; although differences between hospitals are borderline significant and the unexplained variance between patients with regard to Braden scores is much higher.

Conclusion

This chapter is an application-oriented introduction to linear multilevel modeling. It explained the main ideas in multilevel modeling, described analytic procedures, and illustrated interpretation of results. It showed how multilevel modeling allows differentiation between relationships of individual characteristics (lower-level characteristics), group characteristics (higher-level characteristics), and the outcome of interest in a single analysis. It demonstrated how to make inferences regarding the impact of individual characteristics after controlling for groups characteristics or vice versa. The chapter has also shown how multilevel modeling can assess whether the relationship between individual characteristics and the outcome differs in various groups. Finally, advantages of multilevel modeling compared to fixed effects modeling were summarized: (a) it accounts for nested data structures; (b) it allows testing of covariate effects at the appropriate level of analysis; (c) it avoids misleading results associated with the ecological fallacy and inflation of type 1 error rate; (d) unbalanced data may be used; (e) differentiated insights about associations from different levels on the outcome are obtained; and (f) it is easy to conduct and interpret. For further reading we recommend Raudenbush and Bryk (2002), Verbeke and Molenberghs (2000), or Zuur et al. (2009). The webpage of the Centre for Multilevel Modelling at the University of Bristol contains helpful material (www.bristol.ac.uk/cmm/).

References

- Bergstrom, N., Braden, B. J., Laguzza, A., & Holman, V. (1987). The Braden scale for predicting pressure sore risk. *Nursing Research*, 36, 205–210.
- Bijleveld, C.C.J.H., & van der Kamp L.J.T. (Eds.). (1998). *Longitudinal data analysis. Designs, models and methods*. London: Sage.
- Diez-Roux, A. V. (2000). Multilevel analysis in public health research. *Annual Review of Public Health*, 21, 171–192.
- Diez-Roux, A. V., & Aiello, A. E. (2005). Multilevel analysis of infectious diseases. *Journal of Infectious Diseases*, 191(Suppl. 1), S25–S33.
- Drake, D. J., Swanson, M., Baker, G., Pokorny, M., Rose, M. A., Clark-Reed, L., . . . Engelke, M. K. (2010). The association of BMI and Braden total score on the occurrence of pressure ulcers. *Journal of Wound Ostomy Continence Nursing*, 37, 367–371.
- Dunson, D. (Ed.). (2008). *Random effect and latent variable model selection*. New York, NY: Springer.
- Freedman, D. A. (2004). The ecological fallacy. In M. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *Encyclopedia of social science research methods* (pp. 293–295). Thousand Oaks, CA: Sage.
- Hox, J. (2002). *Multilevel analysis. Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.
- Inagami, S., Cohen, D. A., Brown, A. F., & Asch, S. M. (2009). Body mass index, neighborhood fast food and restaurant concentration and car ownership. *Journal of Urban Health*, 86, 683–695.
- Koren, A., & Mawn, B. (2010). The context of unintended pregnancy among married women in the USA. *Journal of Family Planning and Reproductive Health Care*, 36, 150–158.
- Kothari, A. R., & Birch, S. (2004). Multilevel health promotion research: Conceptual and analytical considerations. *Canadian Journal of Nursing Research*, 36, 56–75.
- Kottner, J., Wilborn, D., Dassen, T., & Lahmann, N. (2009). The trend of pressure ulcer prevalence rates in German hospitals: Results of seven cross-sectional studies. *Journal of Tissue Viability*, 18(2), 36–46.
- Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.
- Lahmann, N. A., Halfens, R. J., & Dassen, T. (2005). Prevalence of pressure ulcers in Germany. *Journal of Clinical Nursing*, 14, 165–172.
- Lehmann, E. L. (2005). *Testing statistical hypotheses* (3rd ed.). New York, NY: Springer.
- Maas, C.J.M., & Hox, J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, 46, 427–440.
- Matthews, S. A., & Yang, T. C. (2010). Exploring the role of the built and social neighborhood environment in moderating stress and health. *Annals of Behavioral Medicine*, 39, 170–183.
- Montgomery, D. C., & Peck, E. A. (1992). *Introduction to linear regression analysis*. New York, NY: Wiley.
- Norman, C. D., Maley, O., Li, X., & Skinner, H. A. (2008). Using the internet to assist smoking prevention and cessation in schools: A randomized, controlled trial. *Health Psychology*, 27, 799–810.
- Petree, R. D., Broome, K. M., & Bennett, J. B. (2012). Exploring and reducing stress in young restaurant workers: Results of a randomized field trial. *American Journal of Health Promotion*, 26, 217–224.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R-Development-Core-Team (2011). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1–102 [Computer software].
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods*. London: Sage.
- R Development Core Team (2011). *R: A language and environment for statistical computing*. Retrieved from www.R-project.org
- SAS System for Windows (2008). 9.2 ed. [Computer software]. Cary, NC: SAS Institute.
- Snijders, T.A.B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London: Sage.
- SPSS for Windows (2011). 20.0.0 ed. [Computer software]. Chicago, IL: SPSS.
- StataCorp. (2007). Stata statistical software. Stata/IC 10.1 ed. [Computer software]. College Station, TX: StataCorp.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. New York, NY: Springer.
- Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer.
- Victor, R. G., Ravenell, J. E., Freeman, A., Bhat, D. G., Storm, J. S., Shafiq, M., . . . Leonard, D. (2009). A barber-based intervention for hypertension in African-American men: Design of a group randomized trial. *American Heart Journal*, 157, 30–36.
- Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (Eds.). (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer.

APPENDIX

Syntax commands for the examples in STATA, R, and SPSS. In R we used the following libraries: *nlme* (Pinheiro et al., 2011).

Multilevel linear regression: The intercept-only model

In STATA :

```
xtmixed Braden || hospital:, var
```

In R:

```
library(nlme)
L1 <- lme(fixed=Braden~1, random= ~1 | hospital, data=brad)
summary (L1)
```

In SPSS:

```
MIXED Braden
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT
/RANDOM = INTERCEPT | SUBJECT(hospital) .
```

Multilevel linear regression: Random intercept model with one individual level covariate

In STATA:

```
xtmixed Braden age|| hospital:, var
```

In R:

```
library(nlme)
L2 <- lme(fixed=Braden ~ 1+age , random = ~1 | hospital,
data=brad)
summary(L2)
```

In SPSS:

```
MIXED Braden WITH age
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT age
/RANDOM = INTERCEPT | SUBJECT(hospital) .
```

Multilevel linear regression: Random intercept and slope model with one individual-level covariate

In STATA:

```
xtmixed Braden age|| hospital: age, var
```

In R:

```
library(nlme)
```

```
L3 <- lme(fixed=Braden ~ 1+age, random=~age|hospital, data=brad)
Summary(L3)
```

In SPSS:

```
MIXED Braden WITH age
/PRINT = SOLUTION TESTCOV
/FIXED = INTERCEPT age
/RANDOM = INTERCEPT age| SUBJECT(hospital).
```

11

ANALYSIS OF DYNAMICAL SYSTEMS

The modeling of change and variability

Pascal R. Deboeck and Steven M. Boker

Dynamical systems analysis is characterized by an interest in understanding the change and variability of one or more constructs. While many methods in statistics are focused on the analysis of mean differences, the questions posed in the dynamical systems literature are often focused on differences in the amount and type of variability that occurs in a variable of interest and what variable or variables constitute a regulatory mechanism. In many health-related constructs, the distinctive features may not be in mean differences between groups of individuals but rather in the fluctuations that comprise intraindividual variability.

This chapter begins by identifying scientific problems that can be studied using dynamical systems and defining key terms. A common representation of dynamical systems using derivatives and differential equation modeling will be introduced, followed by a discussion of how one can translate theories about change into testable differential equation models. Finally, approaches to fitting differential equation models are described and resources are identified.

Why dynamical systems?

Many methods in statistics are focused on the analysis of mean differences. Methods such as the analysis of variance are frequently used to draw conclusions about the differences in the means of one or more groups. Applications of regression are used to model how the mean values of a dependent variable differ as a function of one or more predictors. Such methods leave researchers familiar with, and well equipped, to ask mean-difference questions. Mean-difference questions, however, represent only a narrow range of the questions that can be posed. Focusing only on these questions can overlook important questions concerning change and variability.

The distinctive feature of many health-related constructs may not be in mean differences between groups of individuals, but rather in the fluctuations that comprise intraindividual variability: the ebb and flow of depressive or anxiety symptomatology; the waxing and waning of chronic pain; and transitions between confusion and clarity in progressive dementia. These seemingly complex changes reflect dynamics that are not likely to be adequately understood if one considers only mean differences. In many of these cases it is not only desirable to try to understand, and eventually control, mean levels of the construct but also the variation around mean levels. Two patients with similar mean levels of depression might exhibit different etiologies and treatment responses – one patient may have relatively minor variations in affect while the

other shows dramatic variation between extreme states. In many cases, fluctuations around an individual's mean level of a variable may be as important, or even more important, than the value of the mean level itself.

When a set of variables examined together represent the primary inputs and outputs to some quantity, this is called a *system*. When systems exhibit change over time they can be described as *dynamical systems*. Systems that self-regulate or are regulated by an external variable are said to exhibit *deterministic dynamics*. Using statistical techniques to estimate regulatory parameters that characterize such systems is called *dynamical systems analysis*.

Methods that fall under the framework of dynamical systems analysis are characterized by an interest in understanding the change and variability of one or more constructs and *why* they change or regulate in the way they do. Dynamical systems analyses offer a qualitatively different set of questions than mean difference questions, and their application requires learning to ask a different set of questions from one's data. Questions posed in the dynamical systems literature are often more focused on differences in the amount and type of variability that occurs in a construct of interest and what variable or variables constitute a regulatory mechanism.

Understanding regulatory mechanisms is extraordinarily important in clinical sciences, since identification of regulatory parameters can potentially be mapped directly to treatments. For instance, consider the ovulatory cycle. Ovarian hormones have a pattern of coregulation that results in complicated fluctuations of estrogen, progesterone, and leutenizing hormones. The mean levels of these hormones over 30 days provides little information. However, daily levels of these hormones can be modeled as a dynamical system (Boker, Neale, & Klump, 2014) that gives a great deal of information about how the hormones affect each other and how they may be coupled to other fluctuating phenomena such as eating behavior (Hu, Boker, Neale, & Klump, 2013; Klump et al., 2013). Dynamical systems analyses provide the promise of discovery of etiologies of dysregulation as well as the development of individualized timing-appropriate treatments that could not be identified using models for mean differences. By timing-appropriate treatment, we are suggesting that levels of medication or therapy may be most effective if they fluctuate in such a way that the dysregulated system can reestablish its normal regulatory cycle. A mean level of medication may be ineffective whereas a timing-appropriate dosing schedule that results in an appropriately fluctuating level of medication may be therapeutic for diseases of dysregulation. Only by studying dynamic regulating mechanisms can we come to an understanding of how such treatments can be maximally effective.

Figure 11.1 shows depression measured repeatedly over time for two individuals, where each individual is representative of a different treatment group. Typical questions emerging from

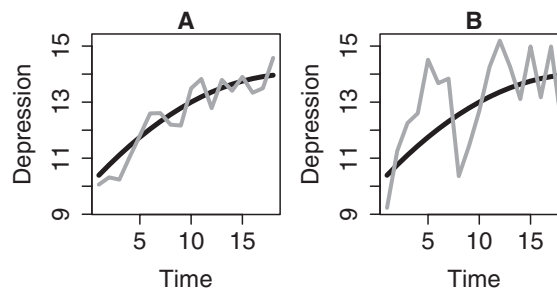


Figure 11.1 Variability around growth trajectories. Observed growth trajectories are shown in gray; modeled trajectories are shown in black. Panels A and B are examples of two individuals, each representative of a differing treatment group, with the same growth trajectories but very different characteristics in how each varies around his/her growth trajectory.

a mean-difference perspective are: (a) Do the groups show a significant difference in means after 18 months of treatment? (b) How does the slope differ for the two groups? (c) Does the slope depend on another construct? In the dynamical systems framework, a very different set of questions is posed, more focused on the changes and variability around each person's respective equilibrium (in this case, the mean trend) (Boker, 2013). Such questions may include: (a) Does one of the groups show more variation around their mean trend? (b) Do there appear to be systematic ways the variation occurs and does this differ between groups? (c) Do changes in depression relate to changes, not just the level, of concurrently measured variables? These two sets of questions are qualitatively different in their foci, and within any given context necessitate the consideration of whether the key questions of interest should be related to mean differences, differences in intraindividual change and variation, or both.

In the context of clinical research, especially when chronicity requires that ongoing care and treatment decisions be made and evaluated, questions that can be answered using dynamical systems analyses may be of significant importance. Too much variation (e.g., in depression, pain, dementia symptoms, hormones, or blood sugar) may be more important to patients and caregivers than the mean level. Reducing large, unexpected fluctuations may allow patients to be more comfortable and support positive adaptation rather than continuing to experience the difficulties and stress presented by dramatic fluctuations. At the other extreme, no fluctuations at all can be symptomatic of dysregulation and may constitute a diagnosable condition (e.g., amenorrhea).

Understanding the antecedents of symptom variation in a patient with dementia (e.g., Lee, Algate, & McConnell, 2013), the changes in anxiety during critical care (e.g., Chlan & Savik, 2011), or establishing good patient-nurse communication feedback (e.g., Happ et al., 2011) first requires questions about change and variation to be addressed. Knowledge about derivatives, differential equation modeling, and methods for fitting these models is needed to understand their usefulness. Citations will be provided for resources that provide more detail and examples of these methods.

Differential equation models

This section begins by introducing *derivatives*, which are mathematical expressions of the way in which one variable changes with respect to another variable. We then present models that utilize derivatives with respect to time, that is *differential equation models*, which are used to express the relationships between the current state of variables and how they are changing with respect to time. We then explore several variations of one common differential equation model (first-order linear differential equation model). Finally, we consider the incorporation of additional variables into differential equation models as predictors or moderators and explain how these modeling decisions reflect differing theoretical perspectives.

Derivatives

Figure 11.2 plots a simulated depression trajectory that might be experienced by a person over the course of 24 months. At any particular time, this person experiences some level of depression, which has been plotted across time (gray line). This person's depression, however, is not characterized solely by level of depression. One could look at the level of depression at several points in time as measurements of the construct depression (shown as the black circles). However, at each point in time, there is also some trajectory suggesting how depression will change in the near future. If over the course of the last month she has become less depressed, one might expect some continuation of this trajectory. Thus at each given moment of time, depression exhibits a *slope*

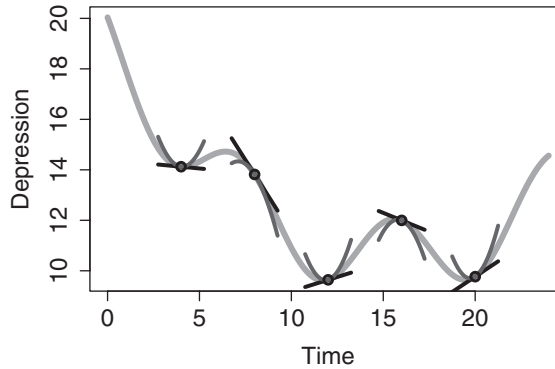


Figure 11.2 Derivatives for a plot of hypothetical depression scores across time (gray line). The instantaneous zeroth derivatives (circles), first derivatives (straight line segments), and second derivatives (curved line segments) are indicated for five different points in time.

(or velocity); that is, the instantaneous rate at which her depression is changing with respect to change in time (shown as the straight line segments). Moreover, the slope of the depression trajectory can be increasing or decreasing across time giving rise to a *curvature* (or acceleration/deceleration) at each moment in time (shown as the curved line segments).

At any given moment in time, the level, slope, and curvature of a variable with respect to time can be expressed mathematically using *derivatives*. In this example we are examining the changes in the construct depression with respect to changes in time. The zeroth derivative x expresses the level of the variable at a specific moment in time in some particular scale, for example points on the Beck Depression Inventory (BDI) (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). In our example, this corresponds to the level of depression at a specific time. The first derivative dx/dt expresses the slope (velocity) of a variable at some specific time, or the limiting value for change in the zeroth derivative dx with respect to the change in time dt as dt approaches 0. In our case, this is the change in depression per specified unit of time. Just like a car's velocity is expressed in "miles per hour," the first derivative of depression could be expressed as "BDI points per day." Similarly, the second derivative d^2x/dt^2 expresses the change in the slope (i.e., curvature or acceleration/deceleration) of a variable at some specific time, or the change in its first derivative with respect to time (change in rate of change). It is important to remember that the second derivative also implies a scale both for the variable and for time, e.g., "BDI points per day per day." "Per day per day" means that each day there is an implied amount of change in the first derivative, which is itself an amount of change per day. The names "zeroth," "first," and "second" refer to what is known as the *order* of the derivative.

Differential equation modeling

Imagining the trajectory of a construct such as that in Figure 11.2 (gray line), the *state* of a construct at any moment in time could be expressed in terms of its current level (a point), its current velocity (a straight line), and its current acceleration (a curved line). That is, the state of a construct at each moment in time can be described using a set of derivatives. A differential equation model can then be used to understand the relations among these derivatives. One could think of a differential equation model as a regression model that uses velocity and acceleration of construct trajectories as well as the levels of constructs. Through these models one can begin to

understand how the states of a construct (level, velocity, acceleration) are related to each other or to the states of other constructs (Deboeck, Nicholson, Bergeman, & Preacher, in press).

A very simple differential equation model posits that the first derivative (rate of change) is constant over time,

$$dx/dt = \beta_1. \quad (11.1)$$

The reader is asked to look on this as they would any other regression equation, imagining that derivatives have been estimated at many points across time and these estimates can be used as one would any other dependent or independent variable. In this equation, values of the first derivative at each point in time are the dependent variables, and there are no independent variables. Later it will be discussed how these equations can be applied to observed data.

Equation 11.1 would be an intercept-only regression model; as with regression, this intercept will correspond to the mean of the dependent variable. In this case, the outcome dx/dt is the velocity (directional speed) at which the construct of interest is changing at every time point, i.e., its slope with respect to time. If β_1 is not significantly different from zero then this would suggest that the mean of the dependent variable, that is the mean of the velocity of the construct (i.e., the slope with respect to time), does not significantly differ from zero.

However, if β_1 differed significantly from zero, this would suggest that the mean velocity (slope of x with respect to time) was greater/less than zero. This would indicate that the average velocity is positive or negative, suggesting an increasing or decreasing trajectory of scores on average. Typically, whether the average velocity (slope with respect to time) of a time series differs from zero is estimated with the regression equation

$$x = \beta_0 + \beta_1(\text{time}). \quad (11.2)$$

The β_1 in Equation 11.1 and β_1 in Equation 11.2 are equal. Equation 11.2 is the result of integrating¹ Equation 11.1. Thus, many dynamical systems models can either be fit in their differential form (Equation 11.1) or in their integrated form (Equation 11.2). There are advantages and disadvantages to each form of the model, but here we will use the differential form since it is more general (not all differential equations have integral solutions) and the integral form becomes much more difficult as the model for change becomes more complex.

First-order linear differential equation model

The strength of differential equation modeling is in the interesting, often seemingly complicated forms of non-linear change that can be expressed when one starts to relate different orders of derivatives to each other. We focus on a first-order linear differential equation, so called because the highest order of derivative that occurs is the first derivative; in this model the first derivative is a linear function of the zeroth derivative (level) at each moment in time. The model

$$dx/dt = \beta x + \varepsilon \quad (11.3)$$

expresses that the velocity of a construct (change with respect to time), dx/dt , is related to the level of the construct x multiplied by β (different from prior β values), plus some distribution of *stochastic innovations* ε . As with other models, β expresses the relation between the independent and dependent variables; in this case β is the relation between the instantaneous level x and velocity dx/dt of a construct at a given time.

One key consideration in such a model is the stochastic innovations, ϵ . These innovations, also called dynamic or process errors, consist of perturbations to the construct that affect its true (unobserved) value. Stochastic innovations are different from measurement error because they do not represent imprecision of scores. Rather, even with perfect measurement constructs may show random (stochastic) perturbations. These innovations are an essential part of the first-order linear differential equation model; changes due to stochastic innovations subsequently would follow the relationship prescribed in the model, unlike measurement error, which is usually assumed to be independent over subsequent observations (see Sijtsma, this volume). The first-order differential equation model will be considered here with three different distributions for ϵ , to each of which independent measurement errors could be added.

No stochastic innovations

We first consider the first-order differential equation with no stochastic innovations, that is $\epsilon = 0$. (For an introduction to deterministic linear differential equations, see Hubbard & West, 1991.) Without stochastic inputs, one can see from Equation 11.3 that when the level, x , is large in magnitude (far from zero), one would expect the velocity, dx/dt , to also be large; whether the velocity would be large in the positive or negative direction would depend on the signs of x and β . Conversely, when x is closer to zero, one would expect the magnitude of the velocity to be smaller. This relationship is pictured in Figure 11.3, where Panels A and B represent positive and negative values of β , respectively, for positive values of x . As can be observed by the dark line segments drawn on this figure (tangent lines), when the x (i.e., score) is closer to zero the velocity is smaller in magnitude (line segments are shallower) and when x (score) is far from zero the velocity is larger (line segments are steeper). The parameter β , can also be used to understand the time required for the level of the dependent variable x to be reduced by half (if negative) or to double (if positive).²

Non-zero stochastic innovations

We next consider non-zero stochastic innovations. Non-zero stochastic innovations (even in the absence of measurement error), and the perturbations are likely to propagate forward in time as the construct continues to change. One common distribution for these innovations are independent, normally distributed observations (Oud & Jansen, 2000; Voelkle, Oud, Davidov, & Schmidt, 2012).³ When this assumption holds, the function in Equation 11.3 exhibits very different behavior as shown in the Panels C and D of Figure 11.3. In this case, the time series that is produced does not consist of a single exponential decay as in the prior case, but instead the construct repeatedly decays towards zero as the innovations perturb the construct above and below a value of zero (other constants can be used). The parameter β now called the *auto-effect*, conveys information about how quickly this decay occurs, and therefore will be related to how correlated scores will be over time. When β is a large negative, values decay quickly (i.e., resulting in a large amount of change dx/dt); therefore, the larger proportion of subsequent observations will be due to new stochastic innovations rather than the prior innovations. Conversely, if β is closer to zero, decay of the current value will be less rapid and the effect of prior innovations will persist for a longer period of time.

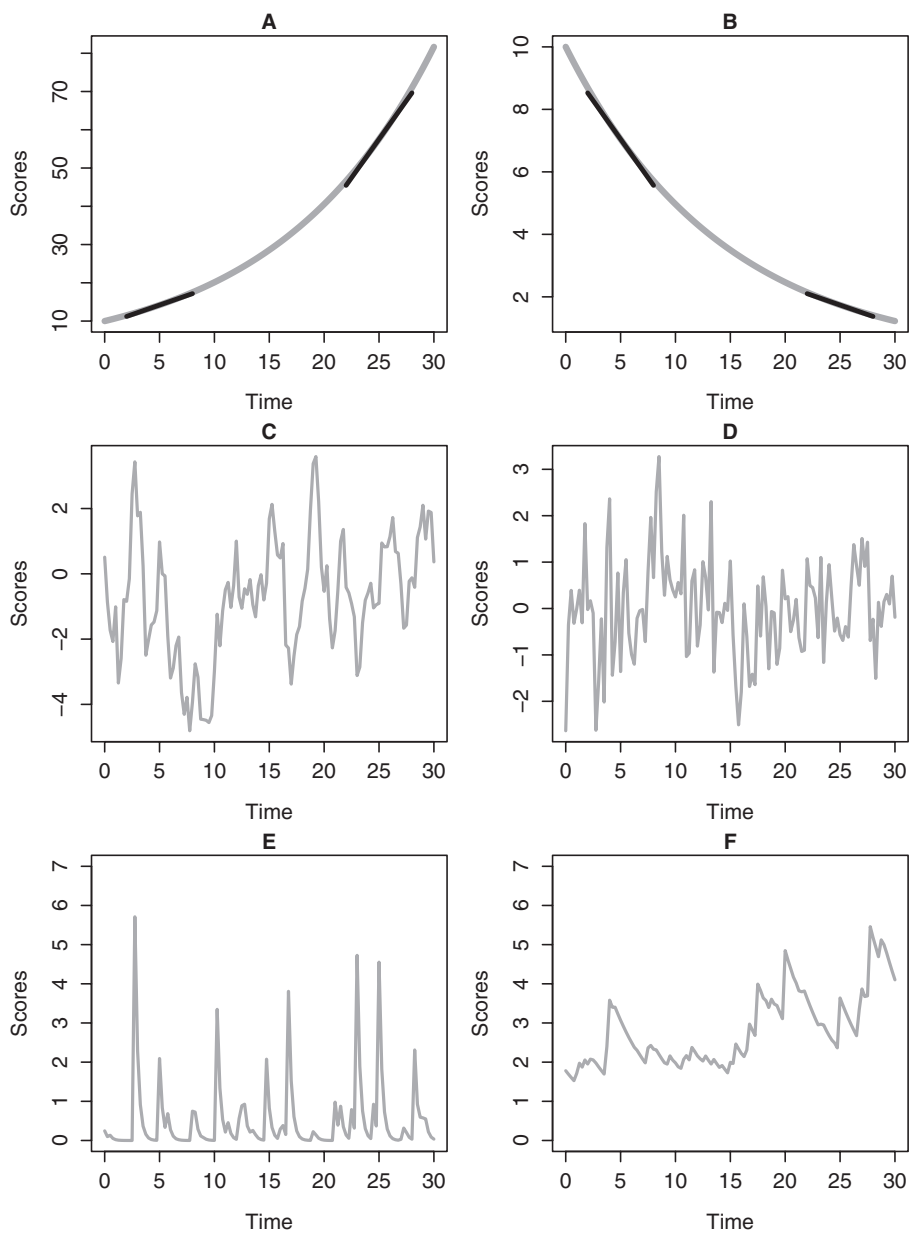


Figure 11.3 Linear model for the first derivative. Three examples of the Equation 11.3 model $dx/dt = \beta x + \varepsilon$. The stochastic innovations ε are equal to zero in Panels A and B, have a normal distribution in Panels C and D, and a distribution consisting only of positive values in Panels E and F. In Panels A and B, β is positive and negative, respectively. In Panels C and D, β is -0.11 and -2.30 , respectively. In Panels E and F, β is -0.92 and -0.05 , respectively.

Positive stochastic innovations

While the stochastic error ϵ is often considered a nuisance, there may be cases where its interpretation may be of substantive interest. When other distributions of ϵ are considered, the behavior of the model in Equation 11.3 changes. For example, consider stochastic innovations consisting of only positive values (Deboeck & Bergeman, 2013). Perhaps these inputs represent the presence of stressors that serve to increase the level of a construct like anxiety. In this case, the model will produce time series such as those in Panels E and F of Figure 11.3. While qualitatively different, this case has similarities to the prior cases in that as the series is perturbed by the stochastic innovations, it is followed by exponential decay in subsequent observations. When stochastic innovations are only positive, the parameter β conveys information that might be considered a rate of dissipation – or how quickly stochastic innovations are depleting over time. This model was created envisioning a reservoir, where the height of the contents of the reservoir corresponds to the level of the construct of interest. The stochastic innovations serve to increase the contents of the reservoir (i.e., increasing level of the construct), and β represents effects of some process used for removal or dissipation of the reservoir contents that decrease the height of the liquid. When the innovations, on average, are of lesser magnitude than the rate of decay, the resulting series looks more like Figure 11.3E. When the rate of decay and innovations are more closely balanced, series such as Figure 11.3F result. In this example, the average of the stochastic error distribution may convey important information regarding a person's perceived inputs on a construct of interest (e.g., Bergeman & Deboeck, 2013).

Translating theory into model

Selecting an appropriate model

The purpose of introducing the first-order differential equation model shown in Equation 11.3 and Figure 11.3 is to highlight that relatively complex forms of change can be described using differential equation modeling, even with a single, relatively simple, model. While this model can be used to produce relatively simple trajectories with no stochastic inputs (Figure 11.3, Panels A and B), the presence of stochastic inputs can result in intricate variation over time (Figure 11.3, Panels C, D, E, and F). Moreover, the substantive interpretation of each of these models lends itself to differing applications.

The first model with no stochastic inputs may be ideal for cases where change in a construct is not expected to have random innovations, but rather progress systematically in some direction. Such a model might be reasonable when assessing longer-term trends in a construct such as chronic pain or spasticity; for example, if a patient was observed every three months and a global assessment of chronic or progressive symptoms was made. The second and third models might be better for modeling changes when assessments are made more frequently and the variation in the construct is expected to be more apparent, such as the recording of severity of symptoms each day or even within day.

Selection of an appropriate model depends on theory and prior research about variation in symptomatology. If one expects random events (triggers) to only increase symptomatology, which then dissipate over time (e.g., an arthritis flare up), this is more akin to the third model with positive stochastic components. In this model, given enough time and few enough stochastic inputs, a person's resting state would consist of few symptoms (i.e., the reservoir would empty); however, this model also allows for the fact that continued stochastic inputs, if balanced with the rate of dissipation, can produce constant up-and-down variations as in Figure 11.3F.

However, if a stable non-zero state for symptomatology is expected, and if stochastic events could cause one to have a better or worse day, this is more akin to the second model that included normally distributed stochastic inputs. In this model, perturbations above and below the steady state occur due to stochastic inputs. While not addressed here, one could also consider whether a patient has the ability to regulate a symptom; that is, whether when the symptom reaches high/low levels the patient will self-regulate, perhaps due to some homeostatic process, which will lead the construct to return to some equilibrium state. In such cases a second-order differential equation model may be informative (Steele & Ferrer, 2011; Hu et al., 2013; Bisconti, Bergeman, & Boker, 2004; Oud & Jansen, 2000).

Adding predictors

Once a model for change has been selected, the next step is inclusion of predictors. Two primary types of predictors can be considered – those measured concurrently with the construct of interest, and those representing stable interindividual differences (traits). Predictors measured concurrently with the construct of interest can be incorporated into Equation 11.3 through an addition such as:

$$dx/dt = \beta_1 x + \beta_2 z + \epsilon, \quad (11.4)$$

where z is the additional predictor. The parameter β_1 in Equation 11.4 has the same interpretation as in Equation 11.3 and the model being fit depends on the selection of the distribution of ϵ . The parameter β_2 then indicates whether z was predictive of the velocity at which x was changing (i.e., dx/dt), taking into account the effect of the level of x . Including a predictor in this way allows for the velocity of the construct x to be more positive or negative than expected based on the construct's relations with itself. An example is shown in Figure 11.4. Panel A consists of a first-order differential equation model (Equation 11.3). Panel B shows the concurrent course for predictor z , which could represent any exogenous variable, but here has been imagined to represent the concentration of some medication in the bloodstream as it is taken at regular intervals. Panel C shows the implementation of Equation 11.4, where the predictor z affects the velocity at which x is changing. It is hypothesized in this example that the effect of z is positive, such that when z is high, a positive value will be added to the rate of change for x . By comparing Panels A and C, it is possible to see that the slope (velocity) for a given value of x is higher when z is high than when z is low.

Alternatively, z may represent interindividual differences (traits) with persons j across occasions i . Such quantities can be used to predict differences in β . Recall that when the stochastic innovations are normally distributed across intervals, the parameter β conveys information about how related, or not, subsequent observations are. Then Equation 11.3 becomes:

$$(dx/dt)_{ij} = \beta_j x_{ij} + \epsilon_{ij} \quad (11.5)$$

$$\beta_j = \beta_0 + \beta_1 z_j.$$

Shown in a single equation:

$$(dx/dt)_{ij} = \beta_j x_{ij} + \beta_1 z_j x_{ij} + \epsilon_{ij}, \quad (11.6)$$

it can be seen that the relationship between x and dx/dt can be allowed to change based on z – that is, the relation between x and dx/dt can be moderated by z (see Coffman, this volume). Figure 11.5 shows an example of β varying as a function of the individual trait z . Panels A, B,

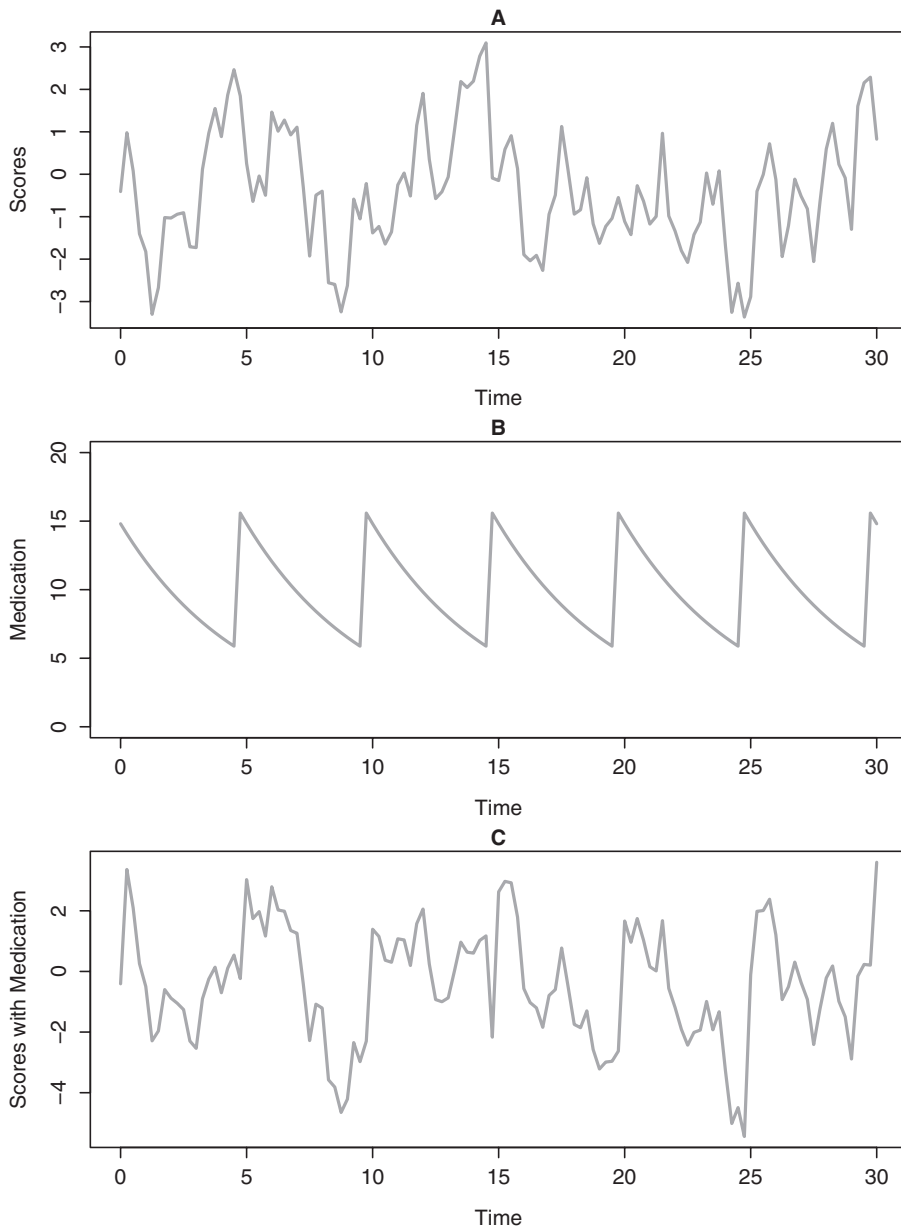


Figure 11.4 Effects of a concurrently measured predictor on an outcome using a simulated first-order differential equation model with normally distributed errors. Panel A shows scores over time without considering the predictor. Panel B shows hypothetical changes in a predictor (e.g., medication blood concentration across repeated administration of a drug on a regular dosing schedule). Panel C shows the additive effects of medication blood concentration on the simulated score trajectory in Panel A. When the medication level is high, a positive change score is added to the simulated data; when the medication level is low, a negative change score is added to the simulated data.

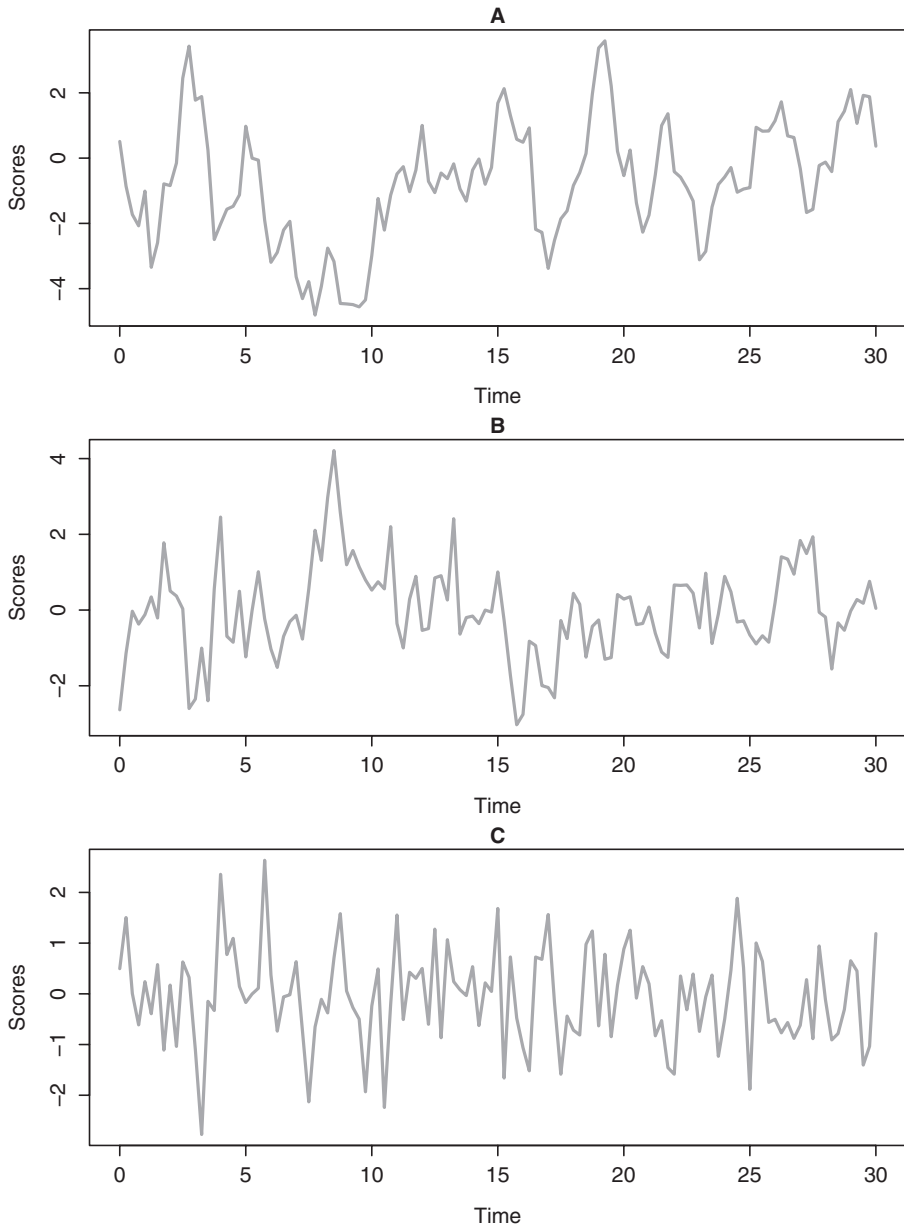


Figure 11.5 Moderating effects of z on β in a first-order differential equation model. Panels A, B, and C show plots of simulated scores resulting when the relation β is correlated with the interindividual trait z for high, moderate, and low values of z , respectively.

and C show individuals high, medium, and low in trait z , respectively. (In this example it has been assumed that z and β are positively correlated such that individuals higher or lower in z will tend to have higher or lower values of β , respectively.) Consequently, how long an innovation persists over the course of the time series (the correlation from one observation to the next) varies depending on the level of the trait z .

These two simple examples constitute the tip of a much larger iceberg. When predictors are time-varying, change in the predictor with respect to time (its derivatives) may be more important than the score (level) at points in time. In some cases the derivatives of predictors may be an important and integral part of models as it is plausible that individuals may be more sensitive to *change* in many contexts rather than the absolute *level* of predictors (Deboeck, Nicholson, Bergeman, & Preacher, 2013). The effects of predictors on the differential equation model, such as the effect of time varying blood concentrations of medication, may also vary as a function of interindividual differences, such as particular metabolic features. The possible range of differential equation models is extensive, but many will test questions about how the level and changes in predictors are related to intraindividual changes, the relationships between one's current level and how one is changing, and the differences in these relationships between different individuals. The range of possibilities may seem to place a burden on researchers to become very specific about the expected relationships between level and change, but the additional flexibility can also be seen as allowing for a more harmonious correspondence between theory and method by not requiring distortion of theory to fit less flexible statistical frameworks. Naturally, the harmony of theory and method is predicated on the ability for appropriate data to be collected. In cases where estimation of individual model parameters are of interest, it is necessary to collect sufficient data from each individual to estimate parameters of the model of interest for each person (typically 30+ observations per person). When groups of people can be assumed to have similar dynamics as few as 3–4 observations per individual may be sufficient (Boker & Nesselroade, 2002).

This section also serves to highlight that differential equation modeling is not a specific statistical methodology, but rather a more general mathematical representation of change. Once a theoretical model has been selected, the application of one or more domains of statistical methodology may be required to test a model. For those familiar with multiple statistical methods, the prior examples suggest models that can draw on prior experience with structural equation modeling (see Schmiede & Bryan, this volume), multilevel modeling (see Grittner & Lahmann, this volume), and possibly the nascent field of multilevel structural equation modeling. Depending on the specific theoretical model one desires to implement, these areas of statistics can serve as valuable background knowledge for allowing more ready application of differential equation model concepts. Readers uninitiated in structural equation modeling or multilevel modeling, and interested in fitting differential equation models, will be well served by the time taken to learn more about these methods.

Fitting differential equation models and resources

The primary challenge in fitting a differential equation model is that only the observed levels (zeroth derivatives) of variables and predictors of interests are typically recorded rather than the level, velocity, and acceleration. Multiple methods have evolved for fitting differential equation models, due in part to the variety of scientific contexts in which differential equation modeling has been applied and the different kinds of data common in those fields. Three approaches that can be used in structural equation modeling (SEM) software are presented here – observed derivative estimates, latent derivative estimates, and the exact discrete model. For each method we provide a brief introduction and identify advantages and disadvantages. Citations are included to serve as resources to learn more; many cited resources include substantive examples and sample code for implementing the method being discussed.

Observed derivative estimates

One approach to estimating derivatives is to calculate observed estimates of the derivatives of a time series. Several methods exist that produce derivative estimates through the linear combination (addition/multiplication) of observed values. These include local linear approximation (LLA; Boker & Nesselroade, 2002; Boker & Laurenceau, 2005), generalized local linear approximation (GLLA; Boker, Deboeck, Edler, & Keel, 2010), and generalized orthogonal derivative estimates (GOLD; Deboeck, 2010, 2011). In these methods a short series of observations (e.g., four sequential observations, which could constitute part of a longer time series from a single individual) is multiplied by a series of weights and summed. When appropriate weights are selected, derivative estimates are produced.

LLA estimates the zeroth derivative as the observed score at any given time, the first derivative as the difference between two observations divided by time, and the second derivative as the change in the rate of change (a differences in slopes) over some period of time. LLA is limited to using three equally spaced observations to estimate derivatives up to the second derivative; GLLA is a generalization of LLA that allows for more than three observations to be used to estimate a derivative, and to allow any order of derivative to be estimated. GOLD is an alternative method for estimating derivatives based on a hierarchical, rather than simultaneous, fitting of polynomials; this method allows for any number of observations to be used to estimate a derivative, any order of derivative to be estimated, and does not require equally spaced observations.

These methods have the advantage of producing observed estimates of derivatives. When these methods are used in conjunction with multilevel modeling or SEM software, the derivative estimates can be used like any observed predictor or outcome. This flexibility allows their usage with a wide variety of statistical software and statistical methodology. These methods, however, require several observations to estimate each set of derivatives; this can be problematic when the true state of the system changes quickly relative to the sampling rate (see Harrison, this volume). Using four observations to estimate derivatives on data collected on a daily basis, for example, will allow observed derivative methods to be well attuned to more gradual changes occurring over half a week, but are likely to do a poor job if rapid shifts occur on a daily basis. This is due to these methods being an approximation relative to other more mathematically exact methods.

GLLA, and the special case LLA, have been applied to a variety of substantive problems. Examples include: rapid cycling bipolar disorder (Boker, Leibenluft, Deboeck, Virk, & Postolache, 2008), resiliency in older adults (Montpetit, Bergeman, Deboeck, Tiberio, & Boker, 2010), rhythmic movements (Butner, Amazeen, & Mulvey, 2005), coping with widowhood (Bisconti, Bergeman, & Boker, 2006), and adolescent substance use (Boker & Graham, 1998). GOLD estimates have not been applied as widely at this time, but an example is included in Deboeck (2011). Selection among these methods depends on the specific differential equation model being fit. When derivatives of the same construct serve as dependent and independent variables and furthermore when derivatives on differing sides of the equal sign are separated by an even number of orders,⁴ GOLD is recommended because correlations in the errors of GLLA estimates can yield biased parameter estimates. For other models, a clear advantage for either method has not been demonstrated at the time of writing.

Latent derivative estimates

Another approach is to estimate the unobserved derivatives that would lead to an observed series of scores; that is, estimates of the latent derivatives of a time series. Like observed derivative methods, the latent derivative (LDE; Boker, Neale, & Rausch, 2004; Boker, 2012) approach uses

a short series of observations (e.g., four sequential observations) to estimate latent derivatives. This method is akin to latent growth curve modeling (McArdle & Epstein, 1987), where latent intercept and slope terms can be estimated from a series of observations. Like latent growth curve modeling, the values of the paths between the latent and observed variables are fixed to identify interpretation of the latent derivatives. Substantive examples of latent differential equation modeling include: coupling of depression and behavior in mothers and children (Nicholson, Deboeck, Farris, Boker, & Borkowski, 2011), coupling of physiological and emotional processes in dyads (Steele & Ferrer, 2011), eating behavior and ovarian hormones (Hu et al., 2013), and daily positive and negative affect (Pettersson, Boker, Watson, Clark, & Tellegen, 2013).

LDE gains all of the advantages of SEM, including ability to separate structural and measurement models, ability to use multiple indicators for the state of a construct at any given time, and ability to solve multiple equations simultaneously amongst others. The cost is that not all SEM programs have the multilevel modeling capabilities needed to consider random effects and equations predicting differences in random effects across clusters (e.g., individuals). OpenMx (Boker et al., 2009) and Mplus (Muthén & Muthén, 2009) allow for multilevel capabilities. The field of multilevel SEM is still relatively new, which can present technical challenges such as difficulty in achieving convergence. As with the observed derivative estimate methods, LDE requires several observations to estimate derivatives and consequently these models can be poor approximations of the underlying differential equation model when the true state of the system changes quickly relative to the sampling rate.

Exact discrete model

Another approach is the exact discrete model (EDM). The exact discrete model, rather than estimating derivatives from a time series, approaches model estimation through the integration of a differential equation model. By integrating over some period of time, it is possible to understand the change that is expected from one observation to the next, based on the selected model. The comparison of the expected change, relative to the changes observed, allow for evaluation as to whether a particular model is reasonable for a given set of data.

The exact discrete model for first- and second-order differential equations, with normally distributed innovations in discrete time, has been derived and implemented in the structural equation modeling software OpenMx (Oud & Jansen, 2000; Oud, 2007; Voelkle et al., 2012). The EDM does not estimate derivatives using a series of observations, and consequently the model parameter estimates are expected to be less biased than either the observed or latent derivative estimation methods when constructs are changing quickly relative to the sampling rate. The equations for the exact discrete model, however, are more challenging to modify and adapt to produce novel models than the previously discussed methods. While modification of models is relatively straightforward with the previous methods, users requiring a model that departs from the set of models for which the exact discrete model has been already derived and/or programmed are likely to find the endeavor very challenging. Substantive examples of exact discrete modeling include: the assessment of bidirectional family relationships (Delsing, Oud, & De Bruyn, 2005) and the relation between antisocial behavior and depressive symptoms (Delsing & Oud, 2008).

Conclusions

We introduced the idea of modeling change and variability. While many of the statistical methods that are being widely used are well suited to asking questions about mean differences between groups, or how mean differences vary as a function of a continuous predictor, the methods we

discussed have the potential to address qualitatively different questions. These questions are about the dynamics of one or more constructs – questions about the changes in constructs and the intraindividual variability of constructs over time. These questions consist of addressing the relations between the current state of one or more constructs and how these constructs are changing. Mathematical expression of these relations, called dynamical systems analysis, can be accomplished using differential equation modeling. Differential equation models, made up of relationships between derivatives, are a precise way to express mathematical theories of change. Learning to use the language of derivatives allows for the expression of theoretical relations between the states of constructs – the level, velocity, and acceleration.

Construction of a theoretical model is probably the most challenging aspect of these methods. After specifying a model, researchers can draw on commonly available statistical methods such as SEM and multilevel modeling. These methods can be used in conjunction with methods for fitting differential equation modelings such as the observed/latent derivative estimates or the exact discrete model, to fit these models. Fundamentally we believe it is important to highlight that the questions that can be addressed by examining change and variability are not better or worse than mean difference questions. Rather that these differing sets of questions necessitate the consideration in a particular scientific context whether the key questions of interest should be related to mean differences, differences in how people change and vary, or both.

Notes

- 1 Integration is the solving of the area under a curve. If one were to plot one's velocity (say in miles per hour) over time (in hours), to find the distance traveled one would need to solve for the area under the curve. If one's velocity were constant (horizontal line), one could solve for the area by multiplying the velocity by the time (solving for the area of a rectangle). When one's velocity is changing, solving for the area under the curve is more challenging, but can be accomplished by multiplying velocity time for very short increments of time (approaching infinitesimally short periods of time) and adding up the distance traveled minute-by-minute. Depictions of *Riemann Sums*, which can be found on the internet, convey the idea of integration.
- 2 The parameter β in this example is called the decay constant in other contexts, such as when studying the decay of radioactive isotopes. In such contexts β can be used to calculate the *half-life*, or the time for the sample to be reduced by half of its original mass. The half-life $t_{1/2}$ is equal to $\ln(2)/\beta$, where $\ln$ is the natural log.
- 3 When expressed as a continuous time stochastic process, this is called the Wiener process, or Brownian motion. When this continuous time process is integrated over some period of time, the probability density function over some increment follows a normal distribution with a mean of zero and a variance that depends on the time over which the process was integrated.
- 4 For example, the first derivative and the fifth derivative are separated by four orders, and the zeroth derivative and the second derivative are separated by two orders.

References

- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, 4, 561–571.
- Bergeman, C. S., & Deboeck, P. R. (2014). Trait stress resistance and dynamic stress dissipation on health and well-being: The reservoir model. *Research on Human Development*, 11, 108–125.
- Bisconti, T. L., Bergeman, C. S., & Boker, S. M. (2004). Emotion regulation in recently bereaved widows: A dynamical systems approach. *Journal of Gerontology: Psychological Sciences*, 59, 158–167.
- Bisconti, T. L., Bergeman, C. S., & Boker, S. M. (2006). Social support as a predictor of variability: An examination of the adjustment trajectories of recent widows. *Psychology and Aging*, 21, 590–599.
- Boker, S., Neale, M., Maes, H., Metah, P., Kenny, S., Bates, T., . . . Spiegel, M. (2009). OpenMx: Multi-purpose software for statistical modeling. University of Virginia, Department of Psychology, Box 400400, Charlottesville, VA 22904. Retrieved from <http://openmx.psyc.virginia.edu>

- Boker, S. M. (2012). Dynamical systems and differential equation models of change. In H. Cooper, A. Panter, P. Camic, R. Gonzalez, D. Long, & K. Sher (Eds.), *APA handbook of research methods in psychology* (pp. 323–333). Washington, DC: American Psychological Association.
- Boker, S. M. (2013). Selection, optimization, compensation, and equilibrium dynamics. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, 26, 61–73.
- Boker, S. M., Deboeck, P. R., Edler, C., & Keel, P. K. (2010). Generalized local linear approximation of derivatives from time series. In S.-M. Chow & E. Ferrar (Eds.), *Statistical methods for modeling human dynamics: An interdisciplinary dialogue* (pp. 161–178). Boca Raton, FL: Taylor & Francis.
- Boker, S. M., & Graham, J. (1998). A dynamical systems analysis of adolescent substance abuse. *Multivariate Behavioral Research*, 33, 479–507.
- Boker, S. M., & Laurenceau, J. P. (2005). Dynamical systems modeling: An application to the regulation of intimacy and disclosure in marriage. In T. A. Walls & J. L. Schafer (Eds.), *Models for intensive longitudinal data* (pp. 195–218). Oxford: Oxford University Press.
- Boker, S. M., Leibenluft, E., Deboeck, P. R., Virk, G., & Postolache, T. T. (2008). Mood oscillations and coupling between mood and weather in patients with rapid cycling bipolar disorder. *International Journal of Child Health and Human Development*, 1, 181–202.
- Boker, S. M., Neale, M. C., & Klump, K. L. (2014). A differential equations model for the ovarian hormone cycle. In P. C. Molenaar, R. Lerner, & K. Newell (Eds.), *Handbook of relational developmental systems: Emerging methods and concepts* (pp. 369–391). New York, NY: Guilford.
- Boker, S. M., Neale, M. C., & Rausch, J. R. (2004). Latent differential equation modeling with multivariate multi-occasion indicators. In K. V. Montfort, J. Oud, & A. Satorra (Eds.), *Recent developments on structural equation models: Theory and applications* (pp. 151–174). Dordrecht, Netherlands: Kluwer Academic.
- Boker, S. M., & Nesselroade, J. R. (2002). A method for modeling the intrinsic dynamics of intraindividual variability: Recovering the parameters of simulated oscillators in multi-wave panel data. *Multivariate Behavioral Research*, 37, 127–160.
- Butner, J., Amazeen, P. G., & Mulvey, G. M. (2005). Multilevel modeling of two cyclical processes: Extending differential structural equation modeling to nonlinear coupled systems. *Psychological Methods*, 10, 159–177.
- Chlan, L., & Savik, K. (2011). Patterns of anxiety in critically ill patients receiving mechanical ventilatory support. *Nursing Research*, 60(Suppl. 3), 50–57.
- Deboeck, P. R. (2010). Estimating dynamical systems, derivative estimation hints from Sir Ronald A. Fisher. *Multivariate Behavioral Research*, 43, 725–745.
- Deboeck, P. R. (2011). Modeling nonlinear dynamics. In M. R. Mehl & T. S. Conner (Eds.), *The handbook of research methods for studying daily life* (pp. 440–458). New York, NY: Guilford.
- Deboeck, P. R., & Bergeman, C. S. (2013). The reservoir model: A differential equation model of psychological capacity. *Psychological Methods*, 18, 237–256.
- Deboeck, P. R., Nicholson, J. S., Bergeman, C. S., & Preacher, K. J. (2013). From modeling long-term growth to short-term fluctuations: Differential equation modeling is the language of change. In R. E. Milsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 427–447). New York, NY: Springer.
- Delsing, M., Oud, J., & De Bruyn, E. (2005). Assessment of bidirectional influences between family relationships and adolescent problem behavior: Discrete vs. continuous time analysis. *European Journal of Psychological Assessment*, 21, 226–231.
- Delsing, M. J. M. H., & Oud, J. H. L. (2008). Analyzing reciprocal relationships by means of the continuous-time autoregressive latent trajectory model. *Statistica Neerlandica*, 62, 58–82.
- Happ, M. B., Garrett, K., Thomas, D. D., Tate, J., George, E., House, M., . . . Sereika, S. (2011). Nurse–patient communication interactions in the intensive care unit. *American Journal of Critical Care*, 19(2), 28–40.
- Hu, Y., Boker, S. M., Neale, M. C., & Klump, K. (2013). Coupled latent differential equations with moderators: Simulation and application. *Psychological Methods*. Advance online publication. doi:10.1037/a0032476
- Hubbard, J. H., & West, B. H. (1991). *Differential equations: A dynamical systems approach*. New York, NY: Springer-Verlag.
- Klump, K. L., Keel, P. K., Kashy, D. A., Racine, S., Burt, S. A., Neale, M., . . . Hu, J. Y. (2013). The interactive effects of estrogen and progesterone on changes in binge eating across the menstrual cycle. *Journal of Abnormal Psychology*, 122, 131–137.
- Lee, K. H., Algase, D. L., & McConnell, E. S. (2013). Daytime observed emotional expressions of people with dementia. *Nursing Research*, 62, 218–225.

- McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, 58, 110–133.
- Montpetit, M. A., Bergeman, C. S., Deboeck, P. R., Tiberio, S. S., & Boker, S. M. (2010). Resilience-as-process: Negative affect, stress, and coupled dynamical systems. *Psychology and Aging*, 25, 631–640.
- Muthén, L. K., & Muthén, B. O. (2009). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
- Nicholson, J. S., Deboeck, P. R., Farris, J. R., Boker, S. M., & Borkowski, J. G. (2011). Maternal depressive symptomatology and child behavior: A dynamical system with simultaneous bi-directional coupling. *Developmental Psychology*, 47, 1312–1323.
- Oud, J. H. L. (2007). Continuous time modeling of reciprocal relationships in the cross-lagged panel design. In S. M. Boker & M. J. Wenger (Eds.), *Data analytic techniques for dynamical systems* (pp. 87–130). Mahwah, NJ: Lawrence Erlbaum.
- Oud, J. H. L., & Jansen, R. A. R. G. (2000). Continuous time state space modeling of panel data by means of SEM. *Psychometrika*, 65, 199–215.
- Pettersson, E., Boker, S. M., Watson, D., Clark, L. A., & Tellegen, A. (2013). Modeling daily variation in the affective circumplex: A dynamical systems approach. *Journal of Research in Personality*, 47, 57–69.
- Steele, J., & Ferrer, E. (2011). Latent differential equation modeling of self-regulatory and co-regulatory affective processes. *Multivariate Behavioral Research*, 46, 994–1003.
- Voelkle, M. C., Oud, J. H. L., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods*, 17, 176–192.

PART IV

Experimental and quasi-experimental design

This page intentionally left blank

12

THEORY-BASED NURSING INTERVENTIONS

Souraya Sidani and Julie Fleury

Nursing interventions are actions that nurses perform with or on behalf of individuals, families, and communities to promote health, support maintenance of function, manage illness, and provide comfort at end of life. Interventions must be carefully designed and implemented to achieve these goals. Interventions should be systematically evaluated prior to use in day-to-day practice. This chapter outlines fundamental issues in intervention research: (a) definition of the term “interventions”; (b) key elements of intervention theory; (c) importance of theory in the design and implementation of interventions and assessment of the effects of interventions on outcomes; and (d) readiness of interventions for translation to practice.

Definition of interventions

Interventions are treatments, therapies, or activities that nurses implement in response to health-related problems to achieve beneficial outcomes (Sidani & Braden, 2011). Nursing interventions arise from the scope of nursing practice (Burns & Grove, 2005), but may be implemented in collaboration with clients and/or other professionals. The Nursing Intervention Classification (McCloskey & Bulechek, 2000) and the Omaha System (2011; see also Westra, Monsen, & Delaney, this volume) provide lists of nursing interventions.

Interventions are rational; they are designed and implemented to address an identified problem and to attain goals that reflect improvements in intended outcomes. Intervention design is based on theory. Theory provides conceptual linkages among the problem requiring remediation or prevention, the intervention, and the expected beneficial outcomes. Intervention theory plays a central role in the design, implementation, and evaluation of nursing interventions and interventions in related health disciplines (Painter, Borba, Hynes, Mays, & Glanz, 2008).

Elements of intervention theory

Intervention theory clarifies conceptual pathways linking problem, intervention, and outcome. It is a melding of explanatory theory of the problem and prescriptive theory of action (National Institutes of Health, National Cancer Institute, 2005). Modifiable aspects of the problem that can serve as intervention targets, processes for inducing change in targeted aspects of the problem, and intervention elements are all part of intervention theory that creates understanding of the what,

for whom, when, why, and how of an intervention. Elements of intervention theory include the problem, the intervention, the outcomes, and mediators and moderators of intervention effects (Foy et al., 2007; Hrisos et al., 2008).

Problems

Problems are health-related conditions of individuals, families, or populations that require improvement, remediation, or prevention (Sidani & Braden, 2011). Problems reflect alterations in biological, cognitive, emotional, behavioral, or social functions, and may signal risk for illness; problems may interfere with performance of healthy behaviors or usual activities, and may ultimately affect well-being and quality of life. Examples of problems are lack of knowledge related to self-management, symptoms such as pain and fatigue, and limitations in physical activity. The North American Nursing Diagnosis Association (NANDA; 2012) and the Omaha System (2011) list problems (diagnoses) that nurses assess and manage. The intervention theory depicts the following aspects of the problem: nature, manifestations, determinants, and level of severity.

Nature of the problem refers to the domain of health in which the problem is experienced. Manifestations are signs and symptoms pointing to the experience or presence of the problem. Determinants are contributing factors. Level of severity is the intensity with which the problem is experienced. In-depth analysis, guided by relevant theoretical and empirical evidence, assists in explaining the inter-relationships among determinants and the problem, thereby specifying pathways through which the problem is caused and sustained (Campbell et al., 2007). Analysis also aids in identifying modifiable aspects that can be targeted to prevent, manage, or resolve the problem.

Interventions

Interventions are a set of inter-related activities directed toward attaining common goals. Intervention theory describes interventions in terms of goals, specific and non-specific elements, mode of delivery, and dose.

Goals

Goals represent the overall direction of the activities comprising the intervention, relative to the targeted problem (Sidani & Braden, 2011). Goals are statements of what the intervention is to achieve. Typical goals are prevention, management, or resolution of a health problem, promotion of health, or enhancement of health-related quality of life.

Specific and non-specific elements

Specific (essential) elements are the “active ingredients” in an intervention. They are actions theoretically expected to induce the changes in the targeted problem and produce beneficial effects on health. Non-specific (non-essential) elements are actions that facilitate implementation of active ingredients but are not anticipated to contribute directly to the outcomes (Stein, Sargent, & Rafaels, 2007). For example, listening to music is an intervention for dyspnea. Tempo of the music is a specific element, whereas encouraging focus on the music is a non-specific element. The specific and non-specific elements are organized into components.

A component is a set of interconnected actions directed toward one goal or targeting one aspect of the problem. Interventions consist of single (simple) or multiple (complex) components.

An example of a simple intervention is acupuncture for prevention of nausea. Diabetes self-management support illustrates a complex intervention targeting multiple factors affecting blood sugar level, such as diet and physical activity.

Mode of delivery

Mode of delivery involves the medium, format, and approach for offering the intervention. Medium is the means by which the intervention is carried out (e.g., written or oral). Format is the specific technique (e.g., booklet or computer-based presentation of written media; face-to-face or videotaped presentation of oral media), and approach is the structure used to provide the intervention. Standardized, or fixed, approaches entail carrying out all intervention activities in a consistent way for all participants. Tailored approaches involve customizing the activities, mode, or dose to be consistent with individual characteristics and needs (Lavori & Dawson, 2008; Lei, Nahum-Shani, Lynch, Oslin, & Murphy, 2012; Murphy, 2005). For example, content of an educational intervention can be matched to the learning needs of individuals so that only topics relevant to identified needs are addressed.

Dose

Dose is the quantity of an intervention, or the level at which an intervention is to be given to produce the expected effects. Dose is a composite that includes amount (e.g., number of sessions, length of session), frequency (number of sessions given within a specified time period), and duration (total time over which the intervention is carried out) (Sidani & Braden, 2011). For example, the dose specification for a fixed behavioral intervention may be: six sessions of 90 minutes each (amount), given once a week (frequency), over a six-week period (duration).

Outcomes

Outcomes are intended consequences of the intervention. They capture change in status following intervention. Although several classifications of outcomes are available (e.g., Doran, 2011; Gershon, Rothrock, Hanrahan, Bass, & Cella, 2010; Moorhead, Johnson, Maas, & Swanson, 2004), the intervention theory specifies immediate and ultimate outcomes guided by the targeted aspects of the problem and the intervention goals. Immediate outcomes are changes in aspects of the problem directly affected by the intervention, occur within a shorter time period, and are operationalized as modifications in the determinants of the problem or alterations in manifestation or level of severity. Ultimate outcomes entail resolution of the problem, prevention of its untoward consequences, or promotion of healthy functioning and quality of life. Typically, ultimate outcomes occur concurrent with or following a series of changes in immediate outcomes (see Liu & Liu, this volume). For example, behavioral interventions designed to increase exercise may target self-efficacy. The intervention creates change in self-efficacy (immediate outcome), which results in increased exercise (ultimate outcome).

Awareness of the nature and inter-relationships among outcomes directs the operationalization of outcomes and specification of the most opportune times for assessing outcomes to determine intervention efficacy and effectiveness (Henly, Wyman, & Findorff, 2011). Clear identification of the specific and non-specific intervention elements and the expected immediate and ultimate outcomes is necessary for elucidating mechanisms underlying intervention effects.

Mediators

The mechanism that mediates the intervention effects reflects the pathway of changes that produce anticipated improvements in the outcomes resulting from the intervention. Mechanism refers to the series of events or alterations in status that occur after receipt of the intervention and that transmit the effects of the intervention to the outcome (MacKinnon, 2008; Nock, 2007; Vallance et al., 2008). In intervention theory, the mechanism is represented by a causal chain linking the intervention to the immediate and ultimate outcome (Coffman, this volume; Suzuki, Yamamoto, & Toshihide, 2011). For more description of mechanisms underlying the effects of interventions, refer to Sidani and Braden (2011), and Fleury and Sidani (2012).

Moderators

Moderators are factors that affect the form or strength of relationship between an intervention and an outcome (Coffman, this volume; MacKinnon, 2008). They may affect the delivery of the intervention such that some moderators may facilitate and others may hinder performance of a subset or all activities of which the intervention is comprised. For instance, the gender composition of clients attending a group session, as well as the gender of the interventionist affect the extent to which clients feel comfortable discussing intimate sexual issues as part of HIV-prevention educational interventions. Moderators may also promote or limit the extent to which the mechanism underlying the intervention effects is triggered. For example, clients with low educational levels or moderate levels of cognitive impairment may have some difficulty understanding all treatment recommendations, thereby precluding them from applying these recommendations and consequently experiencing improvements in outcomes. In addition, moderators can influence clients' responses to the intervention, whereby clients with particular characteristics or with a certain level on such characteristics respond more favorably than others. These favorable responses are operationalized in terms of high improvement in the anticipated outcomes. For instance, clients who led an active lifestyle prior to receipt of cancer therapy may show larger reduction in fatigue following an intervention focusing on physical activity, than those who had a sedentary lifestyle.

Intervention theory identifies moderators of treatment effects. Classes of moderators include characteristics of those who receive the intervention, of interventionists who deliver the treatment, and of the setting in which the intervention is implemented.

Importance of intervention theory

Theory is essential in intervention research. It guides the design, operationalization, implementation, and evaluation of the intervention.

Design

Theory-based design of interventions is systematic, involving four steps: (a) understanding of the problem, (b) identifying modifiable aspects, (c) delineating components, and (d) selecting intervention mode and dose (Sidani & Braden, 2011).

Understanding the problem

Understanding the problem to be treated is the first and fundamental step in intervention design. It involves a review of the conceptual and empirical literature to clarify the problem and identify its manifestations, determinants, and level of severity, as reported in the target population.

Modifiable aspects

The second step focuses on identifying aspects of the problem that are amenable to modification and can serve as intervention targets (Lippke & Ziegelmann, 2008). Propositions from theory guide identification of modifiable aspects of the problem. Judgments about modifiability are made by logical reasoning and critical analysis of each aspect of the problem. For instance, obesity is a problem affecting an increasing proportion of the general population. It is manifested as deposition of fatty tissues in different areas of the body and it results from a variety of factors including genetic predisposition, dietary habits, and lifestyle (i.e., engagement in physical activity). Logic indicates that genetic predisposition is not modifiable (at least up to this date); however, deposition of fatty tissues, dietary habits, and lifestyle are potentially amenable to change: fatty tissues can be removed surgically, whereas dietary habits and lifestyle can be altered if conceptualized as behaviors under the control of individual clients.

Knowledge of modifiable aspects of the problem that are targets for intervention facilitates identification of the population that would benefit from the intervention. People or groups who experience the targeted aspect at the specified level are candidates for intervention. For instance, obese individuals who report having a high-fat diet and engaging in physical activity at a level below recommended are considered for the intervention. In addition, the theory posits external factors related to personal characteristics of clients or features of clients' context that influence the association between the problem and its determinants. These factors may facilitate or hinder the application of the intervention, or moderate its effects on the outcomes. For instance, clients with limited income may not be able to afford purchasing healthy food items (which tend to be expensive) and/or may not have time to engage in structured physical activity (due to other responsibilities or the need to work overtime). The presence of such factors indicates the necessity to design tailored interventions that are consistent with the clients' characteristics, or to adopt a tailored approach for the delivery of interventions. Thus, the theory clarifies the client characteristics forming the basis for tailoring as well as the algorithms for customizing the intervention to be consistent with these characteristics. Alternatively, the factors serve as exclusion criteria (i.e., clients with these characteristics are not given the intervention because they will not benefit from it) or posited and tested as moderators of the intervention effects.

Delineation of therapies

The third step in intervention design entails delineation of therapies (active ingredients) that target modifiable aspects of the problem. The therapies should be congruent with the nature of the modifiable aspects in order to induce the expected outcomes. The therapies are derived from theoretical understanding of the problem and from related empirical literature. For instance, social cognitive theory identifies self-efficacy as a determinant of engagement in healthy behaviors, and indicates four sources of information that if targeted, increase the level of perceived self-efficacy and consequently performance of the behavior. The sources are direct mastery experience, vicarious experience, verbal persuasion, and physiological state. Direct mastery experience involves application of the cognitive skill and behavior whereas vicarious experience is gained by observing others while applying the skill or behavior. Verbal persuasion represents the process of coaching and giving praise to individuals engaging in the behavior. This conceptualization of self-efficacy dictates the incorporation of therapies or strategies into a behavioral intervention that addresses each of these sources if the goal is to increase clients' confidence in carrying out the behavior of interest. The therapies are illustrated with opportunities to practice the behavior in daily life (direct mastery experience), to watch a video-presentation showing the actual

performance of the behavior (vicarious experience), group discussion of challenges encountered when engaging in the behavior and of possible solutions to the challenges (verbal persuasion), and promotion of a supportive context when giving instructions on the behavior performance (to reduce physiological arousal or stress and hence, promote learning and retention).

Selection of mode and dose

The last step in the design of theory-based intervention concerns choosing mode of delivery and dose. The theory proposes the non-specific elements through which the active ingredients are provided (e.g., Thompson, Baranowski, Cullen, & Baranowski, 2007) in a “pure form” (Sidani & Braden, 2011, p. 47). The non-specific elements give directions for specifying the most appropriate medium, format, and approach for delivering the intervention that preserve the purity of the intervention’s active ingredients and therefore, maintain its potency in inducing the mechanism underlying its effects on the ultimate outcomes. For example, therapies for enhancing self-efficacy demand a verbal medium and a group format, which are necessary for demonstrating the correct performance of the behavior and for stimulating group discussion of challenges. Knowledge of the intervention’s active ingredients and mode of delivery suggests how best to integrate these elements into meaningful components and activities and the optimum dose for providing the intervention and for achieving the intended outcomes. For example, the need to offer opportunities to practice the behavior in daily life, with the goal of improving self-efficacy, requires at least two intervention sessions; clients are encouraged to practice the behavior in the time interval between sessions.

Operationalization and implementation

Knowledge of the goals, specific and non-specific elements, delivery mode, and dose direct operationalization and implementation of the intervention. Careful operationalization and implementation are required to maintain theoretical and operational fidelity, and therefore produce the expected outcomes (Borelli et al., 2005). Theoretical fidelity is consistency between the therapies and activities that compose the intervention and the specific elements identified in the intervention theory (Carroll et al., 2007; Keller, Fleury, Sidani, & Ainsworth, 2009). Operational fidelity (integrity) is consistency between the planned and actual delivery of the intervention (Sidani & Braden, 2011).

Theoretical fidelity

Theoretical fidelity arises from a systematic process for deriving intervention components and activities to accurately reflect and be in alignment with its specific and non-specific elements. The process involves translation of the intervention elements into a series of actions, which is facilitated with the generation of a matrix. The matrix links intervention elements with actions to ensure consistency between theoretical and operational elements. For example, direct mastery experience is a specific element of some behavioral interventions. It is operationalized into the following actions: providing information about the behavior, demonstrating the behavior, allowing for reverse demonstration, and providing opportunity for practice. Once specified, the nature and sequence of actions are detailed in an intervention protocol that describes procedures for implementation. The protocol is documented in a manual that clarifies logistics, resources needed, and steps in the intervention. Standards in the intervention manual serve as a reference for assessment of operational fidelity (Sidani & Braden, 2011).

Operational fidelity

Operational fidelity is achieved by implementing the intervention activities consistent with standards in the protocol. Interventionists are often responsible for delivering the intervention and for engaging participants in the application of treatment recommendations or self-management strategies.

Careful selection of interventionists, intensive training of interventionists, and monitoring of their delivery of the intervention contribute to operational fidelity. Interventionists should have the qualifications and personal qualities that enable them to carry out intervention activities. Complete training involves both understanding the theoretical underpinnings of the intervention and mastery of all cognitive and practical skills needed to carry out component activities (Borelli et al., 2005). Training involves explanation and discussion of the intervention theory, as well as careful review of the intervention manual, clarification of the rationale for and the steps of each activity, identification of potential challenges in carrying out each activity and ways to address them, and provision of opportunities for the trainees to practice intervention activities. Optimally, intervention trainees observe an experienced interventionist in action, are supervised when providing the intervention for the first time, and are given constructive feedback, as needed, to ensure their performance meets the standards specified in the manual.

Monitoring implementation is essential (Friedman, Furberg, & DeMets, 1998) to identifying variability or deviation in treatment delivery; in instituting remedial strategies to minimize such deviation; and in obtaining pertinent data for determining the reasons for and the impact of treatment variability on the outcomes. Variability results in differences in enactment of specific elements, at the specified dose; these differences dilute estimates of effects that would otherwise result from the intervention (Carroll et al., 2007; Leventhal & Friedman, 2004). Monitoring intervention implementation rests on specification of component activities and collection of data on the performance of these activities. A list of activities is derived from the intervention protocol and incorporated in a standardized observational or self-report performance assessment (Stein et al., 2007). Assessment of operational fidelity may vary for interventions implemented in different settings, including behavioral (Bellg et al., 2004), psychoeducational interventions (Song, Happ, & Sandelowski, 2010), technology-based interventions (Bosak, Prozehl, & Yates, 2012; DeVito et al., 2011), intensive care units (Chlan, Guttormson, & Savik, 2011), multi-site trials (Robb, Burns, Docherty, & Haase, 2011), and community intervention trials (Breitenstein et al., 2010).

Monitoring of participant implementation of treatment recommendations or self-management strategies is advisable, and has been extensively discussed in terms of compliance or adherence to treatment. Discussion of this topic is beyond the scope of this chapter.

Evaluation of intervention effects

The intervention theory is instrumental in planning and conducting studies aimed to evaluate the effects of interventions on the intended outcomes (Fleury & Sidani, 2012). The explanation of the problem, aspects of the problem targeted by the intervention, and possible moderators of the relationship between the targeted aspects (e.g., determinants) and the problem, helps in specifying the client population and the sample eligibility criteria. The inclusion criteria ensures that clients experiencing the targeted aspects of the problem are included in the evaluation study sample, whereas the exclusion criteria eliminate clients with characteristics that may interfere with the uptake of, engagement in, and adherence to treatment, or influence clients' response to treatment. Awareness of interventionists' characteristics and of contextual factors that may moderate the intervention effects guides the selection of interventionists who possess the qualities posited to

facilitate performance of intervention activities, and the choice of settings that have the physical and socio-cultural features and resources required for appropriate delivery of the intervention. In situations when interventionist and contextual characteristics cannot be controlled, they can be assessed and their influence on outcome achievement can be examined.

Knowledge of the intervention active ingredients or specific elements is necessary for selecting the comparison treatment that serves as control condition for determining the impact of the intervention. The latter treatment should not contain or incorporate any of the intervention specific elements. Information on the intervention specific and non-specific elements, translated into the protocol, gives detailed directions relative to (a) the most opportune time to intervene, that is, the point in time within the trajectory of the problem that is most appropriate to provide the intervention (e.g., before, during, or following experience of the problem); (b) what intervention activities to carry out, how, where, and when over the course of treatment; (c) the sequence for performing the intervention activities and therefore the minimum dose to which clients should be exposed in order to benefit from the intervention; (d) methods for collecting data on intervention dose and for analyzing dose-response relationships; and (e) monitoring the fidelity of intervention implementation and examining its influence on the response to treatment.

Understanding the mechanism creating intervention effects on the anticipated outcomes is required to (a) generate conceptual definitions of mediators and outcomes, which direct the selection or development of appropriate instruments to measure them; (b) delineate the trajectory or pattern of change in the mediators and outcomes, which assists in specifying the points in time during or following treatment for assessing these variables (i.e., when changes in these variables are expected to take place) and the functional form that best describes change (e.g., linear or non-linear) in these variables; and (c) specify the model for testing the proposed relationships among the intervention (or dose), mediators, and outcomes.

Interpretation

As alluded to above, intervention theory indicates what data to obtain and how to analyze the data to determine the extent to which the intervention was effective in producing the intended outcomes. The results of such analysis indicate what exactly contributed to the observed intervention effects and in what way. Results may highlight factors that moderated and mediated relationships among the intervention and the outcomes. In addition, results answer the questions: who would most benefit of the intervention, given by which interventionists, under what context, and what dose, and how the intervention works. Answers to these questions are useful in translating and incorporating interventions in day-to-day nursing practice (Foy et al., 2007).

Readiness for practice

Evaluation of newly designed interventions is done systematically in three main phases. The first phase focuses on examining the feasibility and acceptability of the intervention. Feeley and Cossette (this volume) discuss the design and conduct of pilot studies used to guide refinement of the intervention protocol or theory on which it is based. The second and third phases assess the efficacy and effectiveness of the intervention. Efficacy and effectiveness studies evaluate the effects under ideal and real-world situations, respectively. In efficacy studies, participant, interventionist, and setting characteristics are all carefully controlled, high levels of fidelity are required, and comparison treatment is meticulously selected to maximize the distinction from the active ingredients of the intervention under evaluation; assessment of mediating and outcome variables is conducted at theoretically appropriate times to examine the effects of the intervention in

initiating the mechanism mediating its effects on the outcomes. Wider variation on all factors is typically present in effectiveness trials. Results of these trials indicate the robustness of the effects when the intervention is implemented by different interventionists, to different populations, in different contexts, and point to modifications in mode of delivery and range of dose associated with anticipated outcomes. Interventions with demonstrated effectiveness may be translated into practice.

Summary

Theory-based interventions comprise carefully selected elements translated into components and activities that specifically target and responsively address modifiable aspects of a specific problem. Theory-based interventions target known contributors to a problem. Consequently, theory-based interventions have been found to be more effective in producing the expected changes in outcomes, compared to interventions that are not theory based.

References

- Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., . . . Treatment Fidelity Workgroup of the NIH Behavior Change Consortium. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology, 23*, 443–451.
- Borrelli, B., Sepinwall, D., Ernst, D., Bellg, A. J., Czajkowski, S., Breger, R., . . . Ogedegbe, G. (2005). A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *Journal of Consulting & Clinical Psychology, 73*, 852–860.
- Bosak, K. A., Prozehl, B., & Yates, B. (2012). Challenges of applying a comprehensive model of intervention fidelity. *Western Journal of Nursing Research, 34*, 504–519.
- Breitenstein, S. M., Gross, D., Garvey, C. A., Hill, C., Fogg, L., & Resnick, B. (2010). Implementation fidelity in community-based interventions. *Research in Nursing and Health, 33*, 164–173.
- Burns, N., & Grove, S. K. (2005). *The practice of nursing research. Conduct, critique, and utilization*. Philadelphia, PA: Saunders.
- Campbell, N. C., Murray, E., Darbyshire, J., Emery, J., Farmer, A., Griffiths, F., . . . Kinmonth, A. L. (2007). Designing and evaluating complex interventions to improve health care. *BMJ, 334*, 455–459.
- Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*, 40–48.
- Chlan, L. L., Guttormson, J. L., & Savik, K. (2011). Tailoring a treatment fidelity framework for an intensive care unit clinical trial. *Nursing Research, 60*, 348–353.
- DeVito, D., Song, M.-K., Hawkins, R., Aubrecht, J., Kovach, K., Terhorst, L., . . . Callan, J. (2011). An interventional fidelity framework for technology-based behavioral interventions. *Nursing Research, 60*, 340–347.
- Doran, D. M. (Ed.). (2011). *Nursing outcomes. The state of the science* (2nd ed.). Sudbury, MA: Jones and Bartlett.
- Fleury, J., & Sidani, S. (2012). Using theory to guide intervention studies. In B. M. Melnyk & D. Morrison-Beedy (Eds.), *Designing, conducting, analyzing and funding intervention research. A practical guide for success* (pp. 11–36). New York, NY: Springer.
- Foy, R., Francis, J. J., Johnston, M., Eccles, M., Lecouturier, J., Bamford, G., Grimshaw, J. (2007). The development of a theory-based intervention to promote appropriate disclosure of a diagnosis of dementia. *BMC Health Services Research, 7*, 207–215.
- Friedman, L. M., Furberg, C. D., & DeMets, D. L. (1998). Data collection and quality control. In *Fundamentals of clinical trials* (3rd ed., pp. 156–169). New York, NY: Springer.
- Gershon, R. C., Rothrock, N., Hanrahan, R., Bass, M., & Cella, D. (2010). The use of PROMIS and assessment center to deliver patient-reported outcome measures in clinical research. *Journal of Applied Measurement, 11*, 304–314.
- Henly, S. J., Wyman, J. F., & Findorff, M. J. (2011). Health and illness over time. The trajectory perspective in nursing science. *Nursing Research, 60*(Supp. 3), S5–S14.

- Hrisos, S., Eccles, M., Johnston, M., Francis, J., Kaner, E.F.S., Steen, N., & Grimshaw, J. (2008). Developing the content of two behavioral interventions: Using theory-based interventions to promote GPs management of upper respiratory tract infection without prescribing antibiotics # 1. *BMC Health Services Research*, 8, 11–19.
- Keller, C., Fleury, J., Sidani, S., & Ainsworth, B. (2009). Fidelity to theory in PA intervention research. *Western Journal of Nursing Research*, 31, 289–311.
- Lavori, P. W., & Dawson, R. (2008). Adaptive treatment strategies in chronic disease. *Annual Review of Medicine*, 59, 443–453.
- Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., & Murphy, S. A. (2012). A “SMART” design for building individualized treatment sequences. *Annual Review of Clinical Psychology*, 8, 21–48.
- Leventhal, H., & Friedman, M. A. (2004). Does establishing fidelity of treatment help in understanding treatment efficacy? Comment on Bellg et al. (2004). *Health Psychology*, 23, 452–456.
- Lippke, S., & Ziegelmann, J. P. (2008). Theory-based health behavior change: Developing, testing, and applying theories for evidence-based interventions. *Applied Psychology*, 57, 698–716.
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Lawrence Erlbaum.
- McCloskey, J. C., & Bulechek, G. M. (2000). *Nursing intervention classification* (3rd ed.). St. Louis, MO: Mosby.
- Moorhead, S., Johnson, M., Maas, M., & Swanson, E. (2004). *Nursing outcomes classification (NOC)* (3rd ed.). St. Louis, MO: Mosby.
- Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24, 1455–1481.
- National Institutes of Health, National Cancer Institute (2005). *Theory at a glance. A guide for health promotion practice* (2nd ed.). Retrieved from www.cancer.gov/cancertopics/cancerlibrary/theory.pdf
- Nock, M. K. (2007). Conceptual and design essentials for evaluating mechanisms of change. *Alcoholism. Clinical & Experimental Research*, 31(S3), 4S–12S.
- North American Nursing Diagnosis Association (NANDA). (2012). Glossary of terms. Retrieved from www.nanda.org/DiagnosisDevelopment/DiagnosisSubmission/PreparingYourSubmission/GlossaryofTerms.aspx
- Omaha System. (2011). Problem classification scheme. Retrieved from www.omahasystem.org/problem-classificationscheme.html
- Painter, J. E., Borba, C.P.C., Hynes, M., Mays, D., & Glanz, K. (2008). The use of theory in health behavior research from 2000–2005: A systematic review. *Annals of Behavioral Medicine*, 35, 358–362.
- Robb, S. L., Burns, D. S., Docherty, S. L., & Haase, J. E. (2011). Ensuring treatment fidelity in a multi-site behavioral intervention study: Implementing NIH behavior change consortium recommendations in the SMART trial. *Psycho-Oncology*, 20, 1193–1201.
- Sidani, S., & Braden, C. J. (2011) *Design, evaluation, and translation of nursing interventions*. Ames, IA: Wiley-Blackwell.
- Song, M.-K., Happ, M. B., & Sandelowski, M. (2010). Development of a tool to assess fidelity to a psycho-educational intervention. *Journal of Advanced Nursing*, 66, 673–682.
- Stein, K. F., Sargent, J. T., & Rafaels, N. (2007). Establishing fidelity of the independent variable in nursing clinical trials. *Nursing Research*, 56, 54–62.
- Suzuki, E., Yamamoto, E., & Toshihide, T. (2011). Identification of operation mediation and mechanism in the sufficient-component cause framework. *European Journal of Epidemiology*, 26, 347–357.
- Thompson, D., Baranowski, J., Cullen, K., & Baranowski, T. (2007). Development of a theory-based internet program promoting maintenance of diet and physical activity change to 8-year old African American girls. *Computers & Education*, 48, 446–459.
- Vallance, J. K., Courneya, K. S., Taylor, L. M., Plotnikoff, R. C., & Mackey, J. R. (2008). Development and evaluation of a theory-based physical activity guidebook for breast cancer survivors. *Health Education & Behavior*, 35, 174–189.

PILOT STUDIES FOR RANDOMIZED CLINICAL TRIALS

Nancy Feeley and Sylvie Cossette

A pilot study is a small-scale study conducted to test the plan and method of a research study (Ballman, 2008). A pilot investigation is conducted preparatory to a subsequent, adequately powered study (Conn, Algase, Rawl, Zerwic, & Wyman, 2010) and is designed to try out, evaluate, and select the methods to be used in the larger study (Polit & Beck, 2012). Pilot studies are increasingly important in nursing research because they contribute knowledge about the feasibility and acceptability of research methods and because pilot data are needed to justify use of specific methods proposed in applications for full-scale investigations.

A pilot study may be conducted prior to any type of major study. This chapter will focus on use of pilot studies to inform planning for efficacy trials of nursing interventions. Efficacy trials (Phase III trials) estimate intervention effects under ideal conditions (Campbell et al., 2000) and are commonly called randomized clinical trials (RCTs). The RCT design is a critical tool for creating evidence for practice because it is the most rigorous way to assess causality (Brown, 2002; Rubin & Zell, this volume; Sidani & Braden, 2011). Piloting is particularly important before a full-scale RCT due to their complexities, challenges, and expense. Both the United Kingdom's Medical Research Council guidelines for developing, evaluating, and implementing complex interventions (Craig et al., 2008) and the U.S. National Institutes of Health clinical trial phases adapted for nursing intervention development and testing (Whittemore & Grey, 2002) highlight the importance of testing the protocol of an RCT before proceeding to an evaluation of intervention effects. Thus, the primary purpose of a pilot study is to refine the protocol for the full-scale study by shedding light on strengths, inadequacies, or omissions of the preliminary plan (Conn et al., 2010; Feeley et al., 2009; Loscalzo, 2009; Polit & Beck, 2012; Shanyinde, Pickering, & Weatherall, 2011; Thabane et al., 2010). Findings from the pilot study are utilized to optimize the protocol for the full-scale study, so that it can be successfully and efficiently executed to generate the highest quality evidence for nursing practice.

Pilot study data are not used to estimate intervention efficacy (Polit & Beck, 2012). Reviews of published pilot studies suggest that this is often misunderstood. For example, 81% of "pilot" studies published in seven major medical journals from 2007–2008 inappropriately included hypothesis-testing (Arain, Campbell, Cooper, & Lancaster, 2010). Another review of published pilot study findings indexed in MEDLINE and EMBASE over the past decade found that only 56% addressed methodological issues (Shanyinde et al., 2011). These findings suggest that there is a need to clarify and emphasize what are the appropriate objectives of pilot RCTs.

Aims

The aims of this chapter are to: (a) define feasibility and acceptability, (b) identify design elements that should be assessed for feasibility and acceptability prior to conducting a full-scale trial, (c) describe the decisions that can be taken based on the findings, and (d) summarize current controversies in the use of pilot studies. The focus is on feasibility and acceptability questions emanating from study design. Assessment of feasibility and acceptability of interventions has been described elsewhere (Sidani & Braden, 2011).

Feasibility and acceptability

Both feasibility and acceptability of design elements should be assessed in pilot RCTs. Feasibility is concerned with the researcher's ability to provide the intervention and complete the study as planned (Feeley et al., 2009). For example, the research team's ability to execute the data collection plan should be examined. Acceptability refers to the suitability of the methods or the intervention from the perspective of the study population, the intervention providers, or health care professionals (Feeley et al., 2009). For example, the researcher can explore participants' perceptions and responses to the data collection procedures. Do they find the procedures too time-consuming, inappropriate, or irrelevant? Feasibility and acceptability are often interrelated. If study participants find the data collection too time consuming (e.g., low acceptability), the feasibility of the data collection plan will also be poor.

Feasibility and acceptability of both the study design and the intervention should be examined. The specific feasibility and acceptability questions examined will depend on the particular methodological challenges anticipated in the full-scale trial, as well as challenges related to the provision of the intervention. It is imperative that specific, clear objectives for the pilot study be articulated, and rigorously evaluated. Furthermore, indicators for assessing each study question should also be identified a priori. Explicit criteria should be identified for determining whether to proceed with a specific design element in the full-scale trial (Arain, Campbell, Cooper, & Lancaster, 2010).

Assessing design elements

Overall design

Various study designs can be used in a pilot study preceding an RCT, including experimental or quasi-experimental designs. There are advantages and disadvantages of each of these options. Nonetheless, it is usually optimal to pilot the same design planned for the full-scale study. The number of groups to be included is an important decision. Choices for comparison group(s) include: no intervention (control), usual care, an alternative intervention, or varying doses of the intervention. A three group design is useful to compare the experimental intervention to both an attention control and usual care group, as it controls for the effects of the passage of time and attention. However, this design requires a greater number of participants. In the two group design, the researcher needs to consider which control condition to utilize, and there are limitations to either choice. If the experimental group is compared to usual care any effects observed might be attributed to the extra time and attention provided, whereas when the comparison group is an attention control condition, the intervention effect might not be detectable. To assess the efficacy of different components of complex interventions, factorial designs with two or more crossed components should be used (Maxwell & Delaney, 2003).

In a pilot study, numerous design elements can be examined to inform the design of the full-scale study as virtually every element can be assessed. The researcher decides which elements are most crucial to assess based on their understanding of the inherent methodological and feasibility challenges of the design and procedures. Specific research questions are determined, along with the indicators that will be utilized to answer these questions. Once these pilot data have been collected, decisions are taken based on how to modify the design and procedures for the full-scale RCT that will follow.

For example, in a pilot study of an intervention to reduce mother's anxiety and enhance her ability to interact with her very low birth weight infant in neonatal intensive care (Feeley et al., 2008), one of the research questions was: Is it feasible to observe and measure mother–infant interaction in the neonatal intensive care unit (NICU) prior to intervention? The research staff documented all the challenges and issues arising during their observations of mother–infant interaction, as well as those related to scoring the measure. These data were analyzed and the findings indicated that it was not feasible to measure mother–infant interaction in this context for a number of reasons including not being able to hear mothers' verbalizations, and too little interaction between the mother and her newborn. The decision was made to omit this pre-intervention observation from the full-scale RCT that followed (Feeley et al., 2012). This is just one way in which the pilot study findings shaped the design of the RCT. Furthermore, the findings are also useful to other researchers planning to measure mother–infant interaction in this setting.

Tables 13.1 and 13.2 outline important design elements and procedures that can be assessed for feasibility and acceptability in a pilot prior to an RCT, but this list is not exhaustive. Examples of feasibility and acceptability questions that can guide the pilot study, along with examples of the indicators to answer these questions are included. Decisions concerning the design of the subsequent trial will follow naturally from the findings of the pilot study.

The Consolidated Standards of Reporting RCTs (CONSORT) statement is an evidence-based set of guidelines now widely adopted by journals in many disciplines to improve the reporting of RCTs, and enable readers to evaluate the methodological rigor of these studies (Moher et al., 2010). The CONSORT guidelines can also assist investigators to enhance the design of RCTs and pilot studies by alerting them to the methodological issues that should be addressed (Schulz, Altman, Moher, & CONSORT Group, 2010). At the pilot stage, the researcher should be cognizant of the standards and design the pilot study and the RCT with these guidelines in mind. The following section will discuss some of the key design elements that should be assessed, including those that arise from the CONSORT requirements.

Participant identification and screening

The success of the eventual full-scale RCT depends on enrolling a sufficient number of participants. A well-developed approach to screening potential participants is needed to avoid recruiting and randomizing ineligible persons (Polit & Gillespie, 2010). Thus, developing procedures to identify and assess the eligibility of potential participants should be explored (Conn et al., 2010). Further, the feasibility and acceptability of these procedures should be assessed.

Recruitment and consent

Questions related to recruitment are the focus of a pilot study when the researcher is uncertain about whether sufficient numbers of eligible participants can be accessed at designated study sites. Estimates of the proportion of eligible participants who will agree to participate are

Table 13.1 Design elements and procedures: assessment for feasibility in a pilot RCT

<i>Element/Procedure</i>	<i>Questions</i>	<i>Indicators</i>
Participants	What are the most effective methods for identifying and accessing the population of interest? (e.g., list of appointments, while hospitalized)	• Number of potential participants identified using various methods
Screening	How much time is required to screen potential participants?	• % screened at various sites
Recruitment	What is the number of participants recruited of those screened? (information needed to assess generalizability and provides information about eligibility criteria) Are recruitment strategies effective? What are the characteristics of those who agree to participate? Are these the participants desired? How much time is required to recruit participants? When is the best time to recruit participants? Can an adequate number of participants be recruited at the study sites? (assess the adequacy of settings for recruitment) Can eligible participants provide consent? Is it feasible to provide potential participants with the privacy and time to reflect on their decision?	• Document reasons for exclusion (e.g., type of illness, language barriers, severity of the illness) • % of persons screened who enroll • Characteristics of participants • Length of time to recruit, and time of recruitment
Consent		• Difficulties with consent arising from challenges such as language, literacy, or cognitive ability
Randomization and Blinding	Is the randomization procedure effective? Are the methods used to blind participants, assessors, and/or staff feasible?	• Test various methods to allocate to group (e.g., opaque envelope vs. telephone line vs. computer-generated or website) • Number ineligible persons randomized • Reasons and circumstances for which ineligible persons were randomized
Outcome Measures: Selection	What outcome constructs should be assessed? What measures will best capture the constructs of interest? (i.e., validity)	• Assess two similar constructs and determine which is most appropriate (e.g., a general measure of self-care vs. one that is specific to a particular illness) • Administer two measures of same construct and document responses of participants to each (i.e., ease of responding) • Comments and questions of participants during data collection

Table continues

<i>Element/Procedure</i>	<i>Questions</i>	<i>Indicators</i>
Outcome Measurement: Sensitivity to Change/Timing	Is the outcome measure sensitive to change? When will change occur (possible delayed effects)? How long will change persist (sustained effects)? Are the effects in the expected direction? Are there any negative effects or harms?	• Assess test-retest reliability without any intervention, and it should be high • Assess outcomes at different point of time to identify when change occurs • Seek evidence that intervention has an effect on outcome (eyeballing vs. hypothesis testing, qualitative interviews with participants) • Document and record any unanticipated effects • % measures completed as indicated in protocol • Reasons for deviations from plan • Length of time to complete • Number of missing responses on measure items • Other issues (e.g., missing data in the last questionnaire; sensitive questions at the beginning or the end of the questionnaire) • Inter-rater agreement/reliability • Issues with training of the assessors
Data Collection Procedures	Is it possible to collect data as planned? How long does it take participants to complete the measures? Can participants complete study measures in the time available? Is the sequencing of measures administration appropriate? Are measures being completed by the appropriate participant? (e.g., determine whether peers or family members respond as proxies for participants) Can the outcome assessors be trained to reliably assess the outcome?	• Reasons for withdrawing • Reasons those who are retained remain • % outcome data obtained • Reasons for not obtained outcome data as indicated • Assess strategy to increase retention (e.g., birthday card, regular telephone call)
Retention	What are the rate, timing and reasons for attrition? Is there a relationship between attrition and participant characteristics or site? How effective are the methods of follow-up? What is the best window of time to reach participants for follow-up?	• % of control group participants who are exposed to experimental content • % of participants who cross-over from one group to another
Contamination	Is contamination possible? How?	
Cross-over	Does cross-over of participants from one group to another occur? How?	

Note: The overall design should first be considered in light of the scientific question under study. Indicators should be documented and examined.

<i>Element/Procedure</i>	<i>Questions</i>	<i>Indicators</i>
Recruitment	Will eligible potential participants agree to participate?	• % of persons screened who enroll • Reasons for refusal and how they might be addressed • Difficulties with consent (e.g., language, literacy, or cognitive ability)
Consent	Do eligible potential participants obtain all the information that they desire to make an informed decision?	
Randomization and Blinding	Will participants agree to be randomized? Will participants agree to be blinded? Will clinicians find the allocation methods acceptable?	• Responses to the allocation methods • Reasons for not agreeing to random assignment • Reasons for refusing to be blind to the intervention • Clinicians responses to the allocation methods
Outcome Measures: Selection	Are the measures acceptable to participants? Is one measure more acceptable to participants than another? Can participants comprehend the study measures?	• Assess similar constructs and determine which is most appropriate. (e.g., general versus illness-specific measure of self-care) • Administer two measures of same construct and document responses of participants to each (i.e., ease of responding) • Comments and questions of participants during data collection • Participant responses to systematically queries about the acceptability of measures
Outcome Measurement: Sensitivity to Change/Timing	Is the timing of the outcome assessment acceptable to participants?	• Participants responses to systematic queries about acceptability of the timing of measures
Data Collection Procedures	Will participants agree to and be comfortable providing the data desired? (e.g., blood draws, observations, questions of a sensitive nature) Is the time to complete measures acceptable to participants? Is the sequencing of measures administration acceptable?	• % measures completed as indicated in protocol • Reasons for deviations from plan • Length of time to complete • Missing responses to items • Other issues (e.g., missing data in the last questionnaire; sensitive questions at the beginning or the end of the questionnaire) • Responses to systematic queries about acceptability of procedures
Retention	Will participants find the study procedures acceptable and complete these as planned? What strategies might enhance the retention of participants?	• % outcome data obtained • Responses to interview or questionnaire to elicit reasons for withdrawing, and/or reasons those who are retained remain • Reasons for not obtaining outcome data as indicated • Responses about what strategies enhanced retention
Contamination/Cross-over	Do participants seek other or alternative treatments during the study?	• % of participants who cross-over from one group to another

Note. The overall design should first be considered in light of the scientific question under study. Indicators should be documented and examined.

important for providing evidence of the viability of the RCT in grant applications (Hertzog, 2008), and also provide information about the eventual generalizability of the trial as it may reflect the acceptability of the intervention (Rothwell, 2006). These data should also be used to estimate how much time is required to recruit the desired sample for the full-scale study (Lancaster, Dodd, & Williamson, 2004).

The characteristics of the participants who agree to participate should be examined to determine whether the inclusion and exclusion criteria are effective or require revision (Arnold et al., 2009). Often researchers discover that the inclusion and exclusion criteria as stated for the pilot study require modification because the criteria fail to exclude persons who should be excluded, or are too restrictive and need to be changed to enlarge the pool of eligible persons. As well, the consent process should be explored if the researcher anticipates that this might be challenging. All of these data are used to determine if revisions are needed to the inclusion and exclusion criteria, recruitment procedures or sites, or consent process for the full-scale study.

Randomization and blinding

Randomization is an essential feature of an RCT. If the pilot demonstrates that simple randomization is not suitable to achieve equivalence between groups, other approaches to randomization can be utilized such as block, stratified, or minimization. It is important to ascertain that the target population will agree to be randomly assigned to group (Lackey & Wingate, 1998). There are a variety of reasons that the study population might not be willing to be randomly assigned to group. For example, if the experimental intervention is available, it would be difficult to conduct a trial. People with life-threatening illnesses may not be willing to be randomly assigned to a placebo group if they could have access to cutting-edge experimental treatment.

Various methods can be used to generate the random allocation sequence (e.g., random numbers table, or a computer generated list of random numbers (Moher et al., 2010). A method of concealing the allocation also needs to be utilized, so that staff and others who enroll participants cannot acquire knowledge of group assignment and include or exclude possible participants based on this knowledge (e.g., centralized randomization telephone service or a website (Moher et al., 2010). Feasibility of procedures for generation of the allocation sequence and concealment of the allocation should both be tested at the pilot stage.

Another important methodological feature of RCTs utilized to reduce bias is blinding of participants, their health care providers, outcome assessors, and data analysts (Boutron et al., 2008a). Blinding means concealing information about the group assignment from all of these individuals so that they will not be influenced by this information (Moher et al., 2010). The current CONSORT statement for RCTs of non-pharmacological interventions states that researchers must report the blinding status of participants, their health care providers, and outcome assessors (Boutron et al., 2008b).

Blinding participants in nursing intervention studies is often not possible or difficult to achieve. In many nursing interventions participants often take part in the intervention and are thus fully cognizant of which intervention they receive. Boutron and colleagues (2007) developed a classification scheme to describe creative methods that have been used to blind participants or their health care providers in RCTs assessing non-pharmacological interventions. This classification scheme can be useful for investigators interested in developing and testing different methods of blinding, such as blinding participants to the study hypothesis. Blinding participants to the study hypothesis has been used when participants or their care providers cannot be blinded to the intervention. Participants can be given only partial information in the consent process. Of course, there may be ethical concerns about such methods depending on the context. A modified

Zelen design involving a two-step consent is another approach to blind participants to the study hypothesis (Boutron et al., 2007). As a first step, participants are asked to provide consent to take part in a cohort study (Quilty, Tucker, Campbell, & Dieppe, 2003). After completion of the cohort phases, blinded randomization occurs and only participants randomized to the experimental group are informed that they can receive an experimental intervention. They provide a second consent for this second phase if they wish to do so. The participants allocated to the control group are not informed about the intervention tested in the experimental group and continue to participate in the cohort.

The current CONSORT statement indicates that information concerning how the effectiveness of blinding was assessed is no longer required (Schulz et al., 2010). This is because when researchers assess effectiveness of blinding, most find that it is not successful, and the methods used to assess blinding may not be valid (Sackett, 2007). For example, Hrobjartsson et al.'s (2007) survey of RCTs indexed in the Cochrane Central Register of Controlled RCTs in 2001 found that it was rare for authors to report that they assessed the effectiveness of blinding. Less than half found that blinding was effective, and the most common method used to assess effectiveness was to ask participants to guess their group assignment. Most authors now concur that reporting needs to be improved by describing precisely what was done, to whom, and how (Boutron et al., 2008a; Hrobjartsson, Forfang, Haahr, Als-Nielsen, & Brorson, 2007; Sackett, 2007). Thus, the procedures to assess blinding should be put in place during the pilot study, and their feasibility examined and revised if needed for RCT. However, the assessment of the effectiveness of blinding is currently controversial until more valid methods are developed to do so.

Selection of outcome measures

The pilot study can be used to determine what construct is the best outcome to assess and with what measure (Conn, 2010; Lancaster et al., 2004). For example, the researcher can explore whether a specific or a generic measure of anxiety is more appropriate and sensitive to change. The clarity and acceptability of the measures for participants should also be explored to ensure that the measures chosen capture what the investigators intend them to capture, and that the participants are able to and comfortable providing the data. Test-retest reliability and internal consistency of responses can be assessed. Hertzog (2008) provides guidelines to determine the sample size required to evaluate the psychometric properties of a measure in a pilot study. Information about responsiveness (sensitivity to change) is needed when an outcome will be studied over time (Terwee, Dekker, Wiersinga, Prummel, & Bossuyt, 2003; de Vet & Beurskens, this volume).

Timing of outcome measurement

In nursing intervention studies, the researcher is interested in assessing whether the intervention can bring about the desired change. The anticipated full-scale RCT study design may involve repeated measures, longitudinal follow up, or trajectories as outcomes (Henly, Wyman, & Gaugler, 2011). When making design decisions, the researcher requires an understanding of change processes (Gottlieb & Feeley, 1999). They must understand: When will change occur? How long will it persist? Will there be any delayed effects? In the pilot study, questions concerning the timing of outcome measurements (i.e., duration, frequency), possible mediating variables (Conn et al., 2010), and whether intermediate intervention effects lead to longer-term outcomes, or if short-term outcomes persist (Craig et al., 2008) should be explored if applicable. These data can be particularly useful for justifying every data point in the full-scale study, and their timing. The

issue of whether the effect of the intervention on the outcomes should be examined in a pilot RCT is discussed later in this chapter in the section on current controversies.

Data collection

Many researchers are familiar with the notion that a pilot study should test and refine data collection methods and procedures. A multitude of procedures can be scrutinized including: the time required for data collection, the optimal sequence of measure administration, the feasibility of the data collection (i.e., can the data be collected as planned?), and the acceptability for the participants and clinicians if applicable (i.e., will the participants agree to the data collection procedures?). Assessment of these design aspects is vital to optimize data collection and minimize missing data, thus maintaining power in the full-scale RCT.

The current CONSORT statement for reporting RCTs of non-pharmacological interventions indicates that researchers need to describe any methods used to enhance the quality of their measurements, such as training of outcome assessors. Thus, the pilot study is also an opportunity to address issues, such as inter-rater reliability if applicable.

Participant retention

The timing, rate and reasons for attrition, as well as the relationship of attrition to participant characteristics and site should be studied (Conn et al., 2010). Inclusion and exclusion criteria can then be revised as needed to minimize attrition going forward. Questionnaires can be administered or interviews conducted to determine why pilot study participants drop out, and why they remain.

How effective are the methods for following up study participants? The pilot study also presents an opportunity to implement and assess the effectiveness of evidence-based strategies to retain study participants, in particular when the study design involves repeated measurement of outcomes over time.

Contamination and cross-over

Contamination occurs when participants in the comparison group receive some or all of the components of the experimental intervention (Sidani & Braden, 2011). In nursing intervention studies, participants may be in close proximity to one another, or able to interact, such as in waiting rooms or hospital. They may observe the intervention being provided to others, or they may have opportunities to share information about the intervention. In these ways, participants in the control group may receive the experimental intervention. Cross-over refers to situations where a participant assigned to the control group receives the experimental intervention; or a participant assigned to an intervention group receives instead the control condition or the intervention of another comparison group when more than one intervention is being evaluated (Friedman, Furberg, & DeMets, 2010). The extent of cross-over, reasons for it, and methods for minimizing it should be assessed in pilot studies.

What next? Decision-making following a pilot RCT

How does the researcher decide whether the pilot study findings indicate that methods are sufficiently tested to warrant the full-scale RCT? Unfortunately, there has been little discussion concerning how to arrive at this decision. We propose that prior to the pilot study the researcher

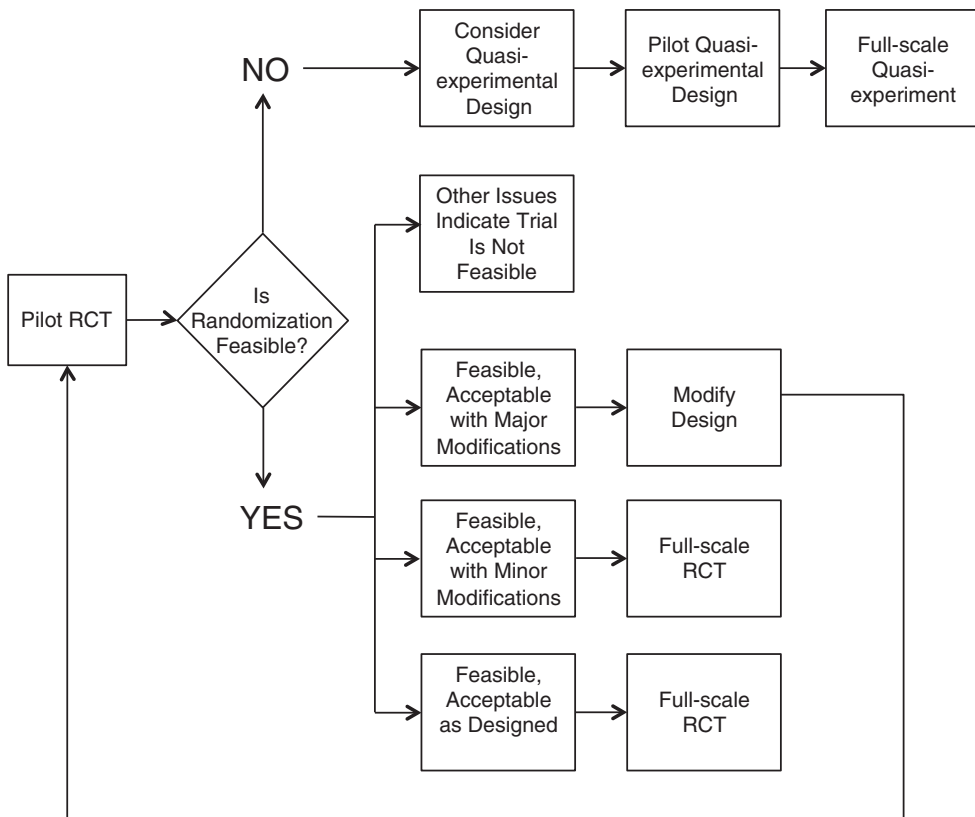


Figure 13.1 Decision-making process from pilot to full-scale RCT.

should identify specific indicators that will be utilized to answer each pilot study question. In Table 13.1 we outline some of the possible indicators for various pilot study questions. For example, if one of the questions is: Will participants find the data collection measures and schedule acceptable? Examples of indicators to answer this question could include: the percentage of measures completed as planned, responses to specific questions about the acceptability of the data collection procedures, and a record of any difficulties encountered. The analyses of the pilot data should then focus on these pre-specified indicators. An overall assessment of these data will guide the ultimate decision about whether to proceed to a full-scale study.

Figure 13.1 describes the various outcomes that might arise following a pilot RCT. If during the course of the pilot it becomes apparent that randomization of participants to group is not feasible or acceptable (i.e., eligible participants do not agree to random assignment or the setting is not suitable for simple randomization), then the researcher needs to consider other designs. These designs may include a cluster randomized trial where a group of patients or different settings (e.g., hospital unit) are randomized instead of individuals (Friedman et al., 2010). A preference trial design can permit participants not willing to be randomized to be included in an RCT. Participants with treatment preferences are given their desired group assignment; while those who do not have strong views are randomly assigned a group as they would be in an RCT (Torgerson & Sibbald, 1998). A historical control design could be employed when the benefit of a new intervention is almost demonstrated, but the effect size has yet to be determined. For

instance, if a new intervention is almost ready to be transferred into clinical practice, a historical control design will permit the researcher to determine the effect of the intervention based on historical data (Friedman et al., 2010).

If randomization is possible, there are four possible outcomes. The first possibility is that the pilot findings indicate that a full-scale RCT is not feasible because of other serious feasibility or acceptability challenges that come to light. Another possibility is that major modifications are required because of problems such as contamination. Another pilot study should be conducted to assess the feasibility and acceptability of the revised study design. The new pilot does not necessarily have to be an RCT if the researcher wishes to assess questions other than random assignment. In many cases, the outcome will be to proceed with the full-scale RCT but with minor modifications to the study design or processes. The last possibility, which is less likely, is that the RCT proceeds but without any modifications as fine-tuning is inevitable.

Current controversies

Sample size for a pilot study

The question of how many participants are needed for a pilot study has received little attention in the literature. As researchers have had little guidance as to how to handle this issue, many pilot studies provide no justification (Shanyinde et al., 2011). Currently, given that the main objective is considered to be an assessment of feasibility and acceptability, a general guideline is that the pilot sample should be large enough to detect flaws in methodology, although factors such as time and budget may also come into play (Thabane et al., 2010). More specifically, several authors propose that the sample size required to answer one of the pilot study's main feasibility questions should be computed using well-known methods to construct confidence intervals around estimates (Arnold et al., 2009; Hertzog, 2008; Thabane et al., 2010). Using this approach, Hertzog (2008) illustrated how to determine sample size based on a pilot study's purpose, and provided guidelines for the sample size required for pilot studies with different purposes.

Estimating effect size in a pilot study

In the past, pilot studies were often used to provide an estimate of effect size prior to proceeding with the full-scale RCT. Conn et al. (2010) proposed that pilot studies should report effect sizes for the outcomes, and these effect sizes should be compared to those in the literature, and discussed in the context of the clinical significance. Because the number of participants in a pilot study is small, estimates of effect size based on these data are imprecise (Hertzog, 2008; Loscalzo, 2009). Thus, some now advocate that estimates obtained from pilot studies should not be used to compute power for the full-scale RCT (Sidani & Braden, 2011; Shanyinde et al., 2011), while others assert that these can be used but very cautiously (Arain, Campbell, Cooper, & Lancaster, 2010; Arnold et al., 2009; Hertzog, 2008; Loscalzo, 2009; Thabane et al., 2010). Hertzog (2008) recommended that researchers estimate confidence limits around a pilot study effect size that has been corrected for bias, and provides information on how to do so.

Most importantly, it is not advisable to draw conclusions about the efficacy of an intervention based on pilot data because sampling variability of the estimates is large. Premature conclusions based on pilot study data could lead to the unwarranted rejection of an intervention that may be efficacious (Loscalzo, 2009; Sidani & Braden, 2011). The opposite is also possible, a significant effect may be observed in the pilot study and not in the full-scale RCT.

Ethical issues

Ethical issues should be considered. CONSORT guidelines specify that any adverse or unintended effects be described in RCT reports (Schulz, Altman, Moher, & CONSORT Group, 2010). The pilot study should explore whether any potential unexpected harmful outcomes arise from the study procedures, such as participant distress. Investigators should also be alert to any indications of collateral harm, such as an increase or decrease in use of needed usual care services due to participation in an intervention study.

In the consent process, participants should be informed that the purpose of the study is to assess feasibility and acceptability questions, and not efficacy (Thabane et al., 2010). The researcher may need to develop an argument to justify the conduct of the pilot study to the Institutional Review Board, indicating that the study purpose is to assess feasibility and acceptability.

Registration and publication

Researchers have proposed that pilot RCTs should be registered in the same way as full-scale clinical RCTs (Arnold et al., 2009; Loscalzo, 2009). Pilot studies are being registered in the database of Current Controlled Trials (De Angelis et al., 2004). Consistent with what is reported in reviews cited earlier, the primary outcome included in these studies labeled as a “pilot study” ranged from feasibility and acceptability issues to a clinical endpoint with statistical testing of the effect of the intervention.

The subject of the publication of pilot studies is highly controversial. Much of the debate on this matter is clouded by the larger question of what is a pilot study, and what are the objectives of such studies. Conn et al. (2010) described the contributions that pilot study reports make to nursing knowledge, including providing information to alert other researchers to problems in methods and procedures and prevent them from unnecessarily conducting a similar study. Specific guidelines for publishing pilot studies based on the CONSORT guidelines have recently been proposed (Thabane et al., 2010). Given the purpose of the pilot study, publications should focus primarily on reporting these findings about feasibility and acceptability (Conn et al., 2010; Thabane et al., 2010).

Conclusion

Pilot studies are an important tool to inform the design and ensure the success of full-scale RCTs of nursing interventions. Thus, it is important that researchers learn to utilize these studies appropriately and effectively to ensure the development of the most rigorous RCTs to contribute to the knowledge needed to guide nursing practice.

Acknowledgment

Nancy Feeley was supported by a Research Scholar Award from the Fonds de recherche du Québec – Santé (FRQ-S).

References

- Arain, M., Campbell, M. J., Cooper, C. L., & Lancaster, G. A. (2010). What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Medical Research Methodology*, 10(67). doi:10.1186/1471-2288-10-67

- Arnold, D.M., Burns, K.E., Adhikari, N.K., Kho, M.E., Meade, M.O., Cook, D.J. for the McMaster Critical Care Interest Group. (2009). The design and interpretation of pilot RCTs in clinical research in critical care. *Critical Care Medicine*, 37(1 Suppl.), S69–S74.
- Ballman, K. (2008, February). *Pilot study design issues*. Presentation at the Center for Health Trajectory Research Seminar Series, School of Nursing, University of Minnesota, Minneapolis. Retrieved from www.nursing.umn.edu/research/research-seminars/SeminarArchive/index.htm
- Boutron, I., Guittet, L., Estellat, C., Moher, D., Hrobjartsson, A., & Ravaud, P. (2007). Reporting methods of blinding in randomized RCTs assessing nonpharmacological treatments. *PLoS Medicine*, 4(2), 0370–0380.
- Boutron, I., Moher, D., Altman, D.G., Schulz, K.F., Ravaud, P., & CONSORT Group. (2008a). Extending the CONSORT statement to randomized RCTs of nonpharmacologic treatment: Explanation and elaboration. *Annals of Internal Medicine*, 148, 295–309.
- Boutron, I., Moher, D., Altman, D.G., Schulz, K.F., Ravaud, P., & CONSORT Group. (2008b). Methods and processes of the CONSORT Group: Example of an extension for RCTs assessing nonpharmacologic treatments. *Annals of Internal Medicine*, 148, W60–W66.
- Brown, S.J. (2002). Nursing intervention studies: A descriptive analysis of issues important to clinicians. *Research in Nursing & Health*, 25, 317–327.
- Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A.L., Sandercock, P., Spiegelhalter, D., & Tyrer, P. (2000). Framework for design and evaluation of complex interventions to improve health. *BMJ*, 321, 694–696.
- Conn, V.S. (2010). Rehearsing for the show: The role of pilot study reports for developing nursing science. *Western Journal of Nursing Research*, 32, 991–993.
- Conn, V.S., Algase, D.L., Rawl, S.M., Zerwic, J.J., & Wyman, J.F. (2010). Publishing pilot intervention work. *Western Journal of Nursing Research*, 32, 994–1010.
- Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: The new Medical Research Council guidance. *BMJ*, 337, a1655.
- De Angelis, C., Drazen, J.M., Frizelle, F.A., Haug, C., Hoey, J., Horton, R., . . . Van Der Weyden, M.B. (2004). Clinical trials registration: A statement from the International Committee of Medical Journal Editors [Editorial]. Retrieved from http://www.icmje.org/clin_trial.pdf
- Feeley, N., Cossette, S., Côté, J., Héon, M., Stremler, R., Martorella, G., & Purden, M. (2009). The importance of piloting an RCT intervention. *Canadian Journal of Nursing Research*, 41, 84–99.
- Feeley, N., Zekowitz, P., Charbonneau, L., Cormier, C., Lacroix, A., Ste. Marie, C., & Papageorgiou, A. (2008). Assessing the feasibility and acceptability of an intervention to reduce anxiety and enhance sensitivity among mothers of very low birth-weight infants. *Advances in Neonatal Care*, 8, 276–284.
- Feeley, N., Zekowitz, P., Shrier, I., Stremler, R., Westreich, R., Dunkley, D., . . . Papageorgiou, A. (2012). Follow up of the Cues and Care randomized controlled trial: Mother and infant outcomes at 6 months. *Journal of Early Intervention*, 34(2), 65–81.
- Friedman, L.M., Furberg, C.D., & DeMets, D.L. (2010). *Fundamentals of clinical trials* (4th ed.). New York, NY: Springer.
- Gottlieb, L.N., & Feeley, N. (1999). Nursing intervention studies: Issues related to change and timing in children and families. *Canadian Journal of Nursing Research*, 30, 193–212.
- Henly, S.J., Wyman, J.F., & Gaugler, J.E. (2011). Health trajectory research: A call to action for nursing science. *Nursing Research*, 60(3 Suppl.), S79–S82.
- Hertzog, M.A. (2008). Considerations in determining sample size for pilot studies. *Research in Nursing & Health*, 31, 180–191.
- Hrobjartsson, A., Forfang, E., Haahr, M.T., Als-Nielsen, B., & Brorson, S. (2007). Blinded RCTs taken to the test: An analysis of randomized clinical RCTs that report tests for the success of blinding. *International Journal of Epidemiology*, 36, 654–663.
- Lackey, N.R., & Wingate, A.L. (1998). The pilot study: One key to research success. In P.J. Brink & M.J. Wood (Eds.), *Advanced design in nursing research* (2nd ed.,) (pp. 375–386). Thousand Oaks, CA: Sage.
- Lancaster, G.A., Dodd, S., & Williamson, P.R. (2004). Design and analysis of pilot studies: Recommendations for good practice. *Journal of Evaluation in Clinical Practice*, 10, 307–312.
- Loscalzo, J. (2009). Pilot RCTs in clinical research: Of what value are they? *Circulation*, 119, 1694–1696.
- Maxwell, S.E., & Delaney, H.D. (2003). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Florence, KY: Taylor & Francis.

- Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., . . . Consolidated Standards of Reporting Trials Groups. (2010). CONSORT 2010 Explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology*, 63(8), e1–37.
- Polit, D. F., & Beck, C. T. (2012). *Nursing research: Generating and assessing evidence for nursing practice* (9th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.
- Polit, D. F., & Gillespie, B. M. (2010). Intention-to-treat in randomized controlled trials: Recommendations for a total trial strategy. *Research in Nursing and Health*, 33, 355–368.
- Quilty, B., Tucker, M., Campbell, R., & Dieppe, P. (2003). Physiotherapy, including quadriceps exercises and patellar taping, for knee osteoarthritis with predominant patello-femoral joint involvement: Randomized controlled trial. *Journal of Rheumatology*, 30, 1311–1317.
- Rothwell, P. M. (2006). Factors that can affect the external validity of randomized controlled trials. *PLoS Clinical Trials*, 1(1), e9.
- Sackett, D. L. (2007). Commentary: Measuring the success of blinding in RCTs: Don't, must, can't or needn't? *International Journal of Epidemiology*, 36, 664–665.
- Schulz, K. F., Altman, D. G., Moher, D., & CONSORT Group. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, c332.
- Shanyinde, M., Pickering, R. M., & Weatherall, M. (2011). Questions asked and answered in pilot and feasibility randomized controlled RCTs. *BMC Medical Research Methodology*, 11(117). doi:10.1186/1471-2288-11-117
- Sidani, S., & Braden, C. J. (2011). *Design, evaluation, and translation of nursing interventions*. Chichester, England: Wiley-Blackwell.
- Terwee, C. B., Dekker, F. W., Wiersinga, W. M., Prummel, M. F., & Bossuyt, P. M. M. (2003). On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Quality of Life Research*, 12, 349–362.
- Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., . . . Goldsmith, C. H. (2010). A tutorial on pilot studies: The what, why and how. *BMC Medical Research Methodology*, 10(1). doi:10.1186/1471-2288-10-1
- Torgerson, D., & Sibbald, B. (1998). Understanding controlled trials: What is a patient preference trial? *British Medical Journal*, 316, 360.
- Whittemore, R., & Grey, M. (2002). The systematic development of nursing interventions. *Journal of Nursing Scholarship*, 34, 115–120.

CAUSALITY IN EXPERIMENTS AND OBSERVATIONAL STUDIES

Donald B. Rubin and Elizabeth R. Zell

Inference for causal effects is extremely important in the clinical sciences and in public health. The field of statistics is critical for addressing such problems using data from both randomized experiments and observational studies. Well-done randomized experiments provide the gold standard for inferring causal effects, whereas drawing such inferences, even in ideal observational studies, requires great care. Here we use the widely utilized Rubin Causal Model (RCM) to define causal effects in both randomized experiments and observational studies, and to draw causal inferences in both settings. The required statistical techniques, although conceptually analogous, can dramatically differ in practice between randomized experiments and observational studies. However, we use well-accepted principles of design and analysis in experiments to bridge to the design and analysis of observational studies, which is the context of many causal inferences in clinical studies.

Randomized experiments

Randomized experiments, often referred to as randomized clinical trials (RCTs) in the clinical sciences, are commonly used to compare an active treatment to a control treatment, often a placebo control. The simplest settings have two groups, where each unit (patient in the health sciences context) has a known probability of assignment into treatment or control, often 0.5. In RCTs, investigators are often blinded to treatment assignment and the units (patients), who are often also blinded to the assigned treatment, are followed for a defined period of time to provide the opportunity to observe outcome variables, generically denoted here by $\mathbf{Y}$.¹ An example is a blood pressure outcome one year after randomization. Randomized trials have strictly developed protocols, with the design of the trial specified in advance of its implementation. Many major journals in the health sciences require that the trial be registered at ClinicalTrials.gov before the trial begins in order for results of the trial to be considered for publication at the conclusion of the study. Many of these journals also require that the registered protocol, and often the statistical analysis plan, be submitted with the manuscript to the journal as part of the peer-review process. Here we denote the observed covariates generically by $\mathbf{X}$, which could include, for example, strata indicators for male versus female, or age category, and family medical history of parents.

A critical feature of randomized trials is that, in expectation, the treated group and the control group are balanced on measured and unmeasured covariates, where *balance* means having the

same expected distributions of all covariates; covariates are variables, like age and baseline blood pressure, thought to be correlated with outcome variables, $\mathbf{Y}$, but which differ from outcome variables because their values are known to be unaffected by the treatment; that is, covariates are known to be the same for each unit whether the unit was assigned to be treated or control.

Units in randomized experiments

In clinical randomized trials involving patients, each unit (i.e., patient) must be informed of the context of the experiment and must give consent prior to enrollment and randomization. This process is often called “informed consent.” Once consent has been given, the unit can be randomized into one of the study groups in the experiment. For simplicity, we consider only two study groups (i.e., *treatment conditions*; an active treatment group and a control group), even though in practice, randomized trials often have more than two groups. The concepts presented here, however, can be easily extended to trials with multiple study groups, or “arms” as they are often called in RCTs.

We start the formal notation for randomized experiments by introducing (not yet defining) the “assignment mechanism.” The definition of assignment mechanism mathematically specifies how treatments are assigned to units. Let $W_i = 1$ if the i th unit ($i = 1, \dots, N$) is assigned to receive the active treatment, and let $W_i = 0$ if the i th unit is assigned to receive the control treatment. In an RCT, the probability that the i th unit is assigned active treatment is between 0 and 1; notationally,

$$0 < P(W_i = 1 | \mathbf{X}_i) < 1, \quad (14.1)$$

where the vertical line indicates conditioning, and $\mathbf{X}_i$ indicates the values of all observed covariates for unit i ; implicitly, the probability in expression (14.1) does not depend on any values of unobserved covariates or on any values of any outcome variables.

Although it is common in randomized trials with two treatment groups to have a 50–50 chance to be assigned to active treatment or to control treatment, that is,

$$P(W_i = 1 | \mathbf{X}_i) = 1/2 \text{ for all } i, \quad (14.2)$$

this is not required. For example, for a disease under study where the outcome “death” is common, units might be more inclined to enroll in the RCT if they believed they had a greater chance of being assigned to the active treatment. In this case the investigators might choose to randomize units to active treatment with probability 2/3 and to control with probability 1/3, no matter what the values of their covariates. Sometimes, this probability can depend on $\mathbf{X}_i$, the observed covariates for unit i ; for example, if $W_i = 1$ indicates assignment to a statin to reduce cholesterol and $W_i = 0$ indicates assignment to a placebo, and X_i indicates baseline cholesterol level, the sicker patients, those patients with larger values of X_i , may be more likely to receive active treatment, by design:

$$P(W_i = 1 | X_i) = \begin{cases} 2/3 & \text{if } X_i > 250 \\ 1/3 & \text{if } X_i \leq 250 \end{cases}. \quad (14.3)$$

Potential outcomes

The causal effect of the active treatment relative to the control treatment for unit i is the comparison of $Y_i(1)$, defined to be the outcome that would be observed when unit i is assigned active

treatment, to $Y_i(0)$, defined to be the outcome that would be observed when unit i is assigned control treatment, with both measured at the same length of time after treatment assignment. It is important to note that in any real-world setting, a unit can only be exposed to one treatment at one point in time, either the active treatment or the control treatment. No unit can be exposed to both treatments at the same point in time. Because we cannot go back in time to give the other treatment, we can only observe $Y_i(1)$ or $Y_i(0)$ for unit i ; thus, the “fundamental problem facing causal inference” (Rubin, 1978) is a missing data problem. Consequently, although causal effects are defined at the level of the individual, they cannot be directly measured.

The collections of observable values are called “the science.” For the N units in the study, the science includes (a) the covariates $\mathbf{X}$, where $\mathbf{X}$ is a matrix with N rows, the i th row being, $\mathbf{X}_i$, (b) the potential outcomes under treatment $\mathbf{Y}(1)$, which is a matrix with N rows, the i th row being $\mathbf{Y}_i(1)$, which gives values for all outcome variables when unit i is exposed to the active treatment, and (c) $\mathbf{Y}(0)$, which is a matrix of the potential outcomes under the control treatment with N rows, the i th row being $\mathbf{Y}_i(0)$ giving the values of the outcome variables for unit i under the control treatment.

For the science, the array $(\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0))$, represents all observable values of Y . We need the stable unit-treatment value assumption (SUTVA), which asserts that each value $Y_i(t)$, $t = 1$ or 0 , is a function only of the unit label i and the assigned treatment t . That is, (a) for unit i , there are no “hidden” treatments not represented by $t = 0$ or $t = 1$, and (b) there is no interference between units; that is, the outcome $Y_i(t)$ for unit i is not affected by the treatments given to any other units. We postpone defining causal estimands, which are comparisons of potential outcomes under the active and control treatments, until after providing a formal and general definition of the assignment mechanism.

The assignment mechanism in general

With notation for the science established, we can now give a formal definition for the assignment mechanism. The assignment mechanism gives the probability of the N -column vector of

treatment assignments $\mathbf{W} = \begin{pmatrix} W_1 \\ \vdots \\ W_N \end{pmatrix}$ given the science:

$$P(\mathbf{W} | \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)). \quad (14.4)$$

This notation reveals the possible dependence of the assignment mechanism on the potential outcomes.

Notice that the expressions (14.1), (14.2), and (14.3) given earlier looked like expression (14.4), but were for randomized experiments and consequently had no dependence on the potential outcomes $\mathbf{Y}(1), \mathbf{Y}(0)$. This is a critical feature of randomized experiments, which are special cases of an *unconfounded assignment mechanism* defined by

$$P(\mathbf{W} | \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)) = P(\mathbf{W} | \mathbf{X}). \quad (14.5)$$

Randomized experiments are also *probabilistic* in the sense that every unit has a positive probability of being assigned either treatment:

$$0 < P(\mathbf{W}_i = 1 | \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)) < 1. \quad (14.6)$$

Moreover, RCTs are “*individualistic*” in the sense that, letting $\mathcal{W}^+$ be the set of $\mathbf{W}$ with positive probability,

$$P(\mathbf{W} | \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)) \propto \prod_{i=1}^N P(W_i | X_i, Y_i(1), Y_i(0)), \text{ for } \mathbf{W} \in \mathcal{W}^+ \quad (14.7)$$

0, for $\mathbf{W} \notin \mathcal{W}^+$.

Any assignment mechanism that is unconfounded (14.5), probabilistic (14.6), and individualistic (14.7) is called *regular*. Regular assignment mechanisms are the bridge between RCTs and observational studies. Any regular assignment mechanism can be written as in (14.7), but with no dependence on the potential outcomes. In other words, in this case, the assignment mechanism is proportional to the product of the N propensity scores, $P(\mathbf{W} | \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)) \propto \prod_{i=1}^N e_i(\mathbf{X})$, for $\mathbf{W} \in \mathcal{W}^+$ where the propensity score e_i for the i th unit is defined as

$$e(X_i) = P(W_i = 1 | X_i). \quad (14.8)$$

In a randomized experiment, the N propensity scores are known, whereas in an observational study they must be estimated – a critical distinction in both design and analysis, as we see later.

Causal estimands

As noted earlier, the fundamental problem of causal inference is one of missing data – there is no purely empirical way to estimate unit-level causal effects. However, typical causal effects can be estimated. For example, a simple estimand compares the average potential outcome under the active treatment with the average potential outcome under the control treatment:

$$\text{ave}_i(Y_i(1)) - \text{ave}_i(Y_i(0)), \quad (14.9)$$

where $\overline{Y(t)}$ is the average value across all N units of the $Y_i(t)$, $t = 1, 0$. Or the estimand could be the median individual causal effect,

$$\text{med}_i[Y_i(1) - Y_i(0)], \quad (14.10)$$

or the difference of medians,

$$\text{med}_i Y_i(1) - \text{med}_i Y_i(0), \quad (14.11)$$

which generally differs from (14.10), whereas (14.9) equals the average individual causal effect $\text{ave}_i(Y_i(1) - Y_i(0))$.

Sometimes the estimands can depend on X_i . Examples are men’s cholesterol reduction when assigned to take a statin versus a placebo, or the woman’s reduction in blood pressure when taking a hypertensive drug versus placebo for those less than 60 years of age. Generally causal estimands are a comparison of $\{Y_i(1) \mid i \in S\}$ values and $\{Y_i(0) \mid i \in S\}$ values where S is an ordered set of indices that can depend on any part of the Science (and even sometimes on $\mathbf{W}$), but commonly depends only on $\mathbf{X}$.

The estimands of interest in a randomized trial should be defined before the trial begins, and should be clearly stated in the protocol and in the statistical analysis plan. The question of

primary interest should be stated clearly in the design phase of the trial: it should not be “constructed” after seeing the final data.

Intuitively, RCTs provide reliable answers to causal questions because we know the rule used to select the treated and control units, and each unit had a known chance to be either. More precisely, consider a completely randomized experiment with half treated and half controls. Then $\{Y_i(1) \mid W_i = 1\}$ are simply a random sample from all $Y_i(1)$ and so fairly represent all $Y_i(1)$; analogously, for the $\{Y_i(0) \mid W_i = 0\}$ and all $Y_i(0)$. Similarly, if the assignment mechanism involves observed covariates, the set $\{Y_i(1) \mid X_i = x, W_i = 1\}$ fairly represents the set $\{Y_i(1) \mid X_i = x\}$ and the set $\{Y_i(0) \mid X_i = x, W_i = 0\}$ fairly represents set $\{Y_i(0) \mid X_i = x\}$, for any value x .

Although this last statement is formally true for any regular design, reliable analysis can nevertheless be tricky in some regular designs when some of the propensity scores are close to zero or one, because then certain values of covariates are far more common in one treatment group than the other group. An extreme example occurs when a covariate value occurs in one observed treatment group but does not occur in the other observed group, making the comparison of $Y_i(1)$ and $Y_i(0)$ values for units with that value of the covariates not immediate.

Observational studies

This last comment points out one of the problems with drawing inferences from observational studies, even those that can be assumed to be regular given covariates $\mathbf{X}$. The distribution of $\mathbf{X}_i$ in the treatment group may be quite different from that distribution in the control group. A regular observational study differs from a RCT because the functional form for the propensity score is unknown in the former. This distinction is important because in a regular design, no matter how many covariates are involved, we have the following fact:

$$\begin{aligned} \text{The set } \{Y_i(t) \mid e(X_i) = e^*, W_i = t\} \text{ fairly represents the} \\ \text{set } \{Y_i(t) \mid e(X_i) = e^*\}, \text{ for } t = 1, 0 \text{ and any values of } 0 < e^* < 1. \end{aligned} \quad (14.12)$$

The design phase of many observational studies assumes the study is regular and tries to create balanced groups on all covariates using (14.12) with estimated propensity $\hat{e}(X_i)$ scores replacing the unknown $e(X_i)$.

Modes of causal inference from observed data

As earlier, assume that $\mathbf{X}$ is fully observed, as is $\mathbf{W}$. We also assume no unintended missing data in potential outcomes, so that if $W_i = t$, $Y_i(t)$ is observed. The observed value of Y for unit i can be written as

$$Y_{i,obs} = W_i Y_i(1) + (1 - W_i) Y_i(0), \quad (14.13)$$

and

$$\mathbf{Y}_{obs} = \begin{pmatrix} Y_{1,obs} \\ \vdots \\ Y_{N,obs} \end{pmatrix}. \quad (14.14)$$

Statistical inference for causal effects uses $\mathbf{X}, \mathbf{W}, \mathbf{Y}_{obs}$ to draw inferences about

$$\mathbf{Y}_{mis} = \begin{pmatrix} Y_{1,mis} \\ \vdots \\ Y_{N,mis} \end{pmatrix}, \quad (14.15)$$

where $Y_{i,mis} = W_i Y_i(0) + (1 - W_i) Y_i(1)$. For notational explicitness, in a particular dataset, we let $\tilde{\mathbf{W}}, \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}_{obs}$ to refer to the observed values of $\mathbf{W}, \mathbf{X}, \mathbf{Y}_{obs}$.

There are two distinct forms of assignment-mechanism-based (or randomization-based) modes of causal inference: one Fisherian and one Neymanian.

Fisherian mode

The Fisherian approach is closely related to the mathematical idea of proof by contradiction. The first element in this mode is the null hypothesis, which is usually $Y_i(1) \equiv Y_i(0)$ for all units: the treatments have absolutely no effect on the potential outcomes. Under this null hypothesis, all potential outcomes are known from the actual observed values of the potential outcomes, $\tilde{\mathbf{Y}}_{obs}$, because $Y_i(1) \equiv Y_i(0)$, and either $Y_i(1)$ or $Y_i(0)$ is observed. Under the null hypothesis, it follows that the value of any scalar statistic, $T(\mathbf{W}, \mathbf{X}, \mathbf{Y}_{obs})$, such as the difference of the observed averages for units exposed to treatment 1 and units exposed to treatment 0, $\bar{y}_1 - \bar{y}_0$, is known, not only for the observed assignment for which its value is $T(\tilde{\mathbf{W}}, \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}_{obs}) = \tilde{T}_{obs}$, but also for all possible assignments $\mathbf{W}$. Suppose we calculate the value of $T(\mathbf{W}, \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}_{obs})$ under each possible assignment W (assuming the null hypothesis) and also calculate the probability of each assignment $W \in \mathcal{W}^+$ under the randomized assignment mechanism. Knowing the value of $T(\mathbf{W}, \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}_{obs})$ for each W and its probability, we can then calculate the probability (under the assignment mechanism and the null hypothesis) that we would observe a value of $T(\mathbf{W}, \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}_{obs})$ as unusual as, or more unusual than, the actual observed value $\tilde{T}_{obs}$. Unusual is defined *a priori*, often by how discrepant $\tilde{T}_{obs}$ is from the typical values of $T(\mathbf{W}, \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}_{obs})$. This probability is the plausibility (p -value or significance level) of the observed value of the statistic under the null hypothesis: if the null hypothesis were true, what is the probability of $T(\mathbf{W}, \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}_{obs})$ being as rare, or more rare, than $\tilde{T}_{obs}$?

Neymanian mode

Neymanism randomization-based inference can be viewed as drawing inferences by evaluating the expectations of statistics, over their distributions induced by the assignment mechanism, in order to calculate a confidence interval for the typical causal effect. First, an unbiased estimator of the causal estimand (the typical causal effect, e.g., the average, $\bar{Y}(1) - \bar{Y}(0)$) is created, and an unbiased, or upwardly biased, estimator of the sampling variance of that unbiased estimator is found; (bias and sampling variance are both defined with respect to the randomization distribution. The typical unbiased estimator in a randomized experiment is the difference in the means of the treatment and control groups, $\bar{y}_1 - \bar{y}_0$, and the estimator of its sampling variance is $s_1^2/n_1 + s_0^2/n_0$, where s_t^2 and n_t are the sample variance and sample size, respectively, in treatment group $t = 1, 0$). Then, an appeal is made to the central limit theorem for the normality of the estimator over its randomization distribution, whence a large sample confidence interval for the causal estimand is obtained.

Nonrandomized studies

With a data set that is not from a randomized study, we try to structure the problem so that we can conceptualize the data as having arisen from an underlying regular assignment mechanism, and then estimate causal effects with the help of estimates of the propensity scores for all the units. A key idea is that, like good experiments, good observational studies are designed, not simply found. As stated earlier, when designing an experiment, we do not have any outcome data, but we plan the collection, organization, and analysis of the data to improve our chances of obtaining valid, reliable, and precise causal answers. The same exercise should be conducted in an observational study: even if outcome data are available at the design stage, they should be set aside. Because observational studies are rarely known to be unconfounded, we are concerned with the sensitivity of answers to unobserved covariates.

Posterior predictive (Bayesian) inference for causal effects requires a model for the Science, $P(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$. A virtue of the RCM framework is that it separates the Science and a model for it, from what we do to learn about the Science – the assignment mechanism, $P(\mathbf{W} \mid \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$. This approach directly and explicitly confronts the missing potential outcomes, $\mathbf{Y}_{\text{mis}}$, by multiply imputing them. That is, the RCM perspective takes the specification for the assignment mechanism and the specification for the underlying data, and derives the posterior predictive distribution of $\mathbf{Y}_{\text{mis}}$, that is, the distribution of $\mathbf{Y}_{\text{mis}}$ given all observed values:

$$P(\mathbf{Y}_{\text{mis}} \mid \mathbf{X}, \mathbf{Y}_{\text{obs}}, \mathbf{W}). \quad (14.16)$$

From this distribution and the observed values of the potential outcomes, the observed assignments, and observed covariates, the posterior distribution of any causal effect can, in principle, be obtained. This conclusion is immediate if we view the posterior predictive distribution (16) as specifying how to take a random draw of $\mathbf{Y}_{\text{mis}}$. Once a value of $\mathbf{Y}_{\text{mis}}$ is drawn, any causal effect can be directly calculated from the drawn value of $\mathbf{Y}_{\text{mis}}$ and the observed values of $\mathbf{W}$, $\mathbf{X}$, and $\mathbf{Y}_{\text{obs}}$. Repeatedly drawing values of $\mathbf{Y}_{\text{mis}}$ and calculating the causal effect for each draw generates the posterior distribution of the desired causal effect. Thus, we can view causal inference entirely as a missing data problem, where we multiply impute the missing potential outcomes to generate a posterior distribution for the causal effects.

An assignment mechanism is ignorable for specific observed values $\tilde{\mathbf{W}}, \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}_{\text{obs}}$, if

$$P(\tilde{\mathbf{W}} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}_{\text{obs}}, \mathbf{Y}_{\text{mis}}) = P(\tilde{\mathbf{W}} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}_{\text{obs}}) \quad (14.17)$$

for all possible $\mathbf{Y}_{\text{mis}}$. All regular assignment mechanisms are ignorable, but some irregular ones that depend on $\mathbf{Y}_{\text{obs}}$, such as sequential designs, are also ignorable. With an ignorable treatment assignment, formally all that we need to model is the science $P(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$, and we can ignore the assignment mechanism. A strength of this model-based approach is that it allows us to conduct causal inference by predicting the missing potential outcomes from observed values. The problem with this approach is the need to specify the distribution $P(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$, which sometimes can implicitly involve extrapolations that are unreliable especially when propensity scores are extreme. More details of this approach are beyond the scope of this chapter. With nonignorable treatment assignment, the simplifications just described do not follow in general, and the analysis typically becomes far more difficult and speculative. Here, sensitivity analyses to missing covariates are important components of satisfactory analyses.

Complications

Many complications may and often do occur in real-world studies for causal effects, many of which can be handled much more flexibly with the Bayesian approach than with randomization-based methods. Of course, the models involved can be difficult to formulate in a practically reliable manner. In addition, Neymanian-style evaluations are still important. Fisherian p -values are a special case of Bayesian posterior predictive p -values, as described in Rubin (1984).

Most of the field of classical experiment design is devoted to issues that arise with more than two treatment conditions and covariates that define complicated blocking structures.

Missing data, due perhaps to unit dropout in large studies, can complicate analyses more than one would expect based on a cursory examination of the problem. Methods such as multiple imputation, the expectation-maximization (EM) algorithm, data augmentation, and the Gibbs sampler are fully compatible with the Bayesian approach to causal inference.

Another common complication is noncompliance. Further complications include partially defined outcomes, such as quality of life that is only fully defined for patients who are alive at the time of measurement. In the real world, complications typically do not appear simply one at a time. For example, an RCT can suffer from missing data in both covariates and longitudinal outcomes, and noncompliance, and partially defined outcomes because of truncation due to death.

Many of the above complications can be viewed as special cases of *principal stratification*. This appears to be an extremely fertile area for research and application of Bayesian methods for causal inferences, especially using modern simulation methods.

Bibliographic notes

The Rubin Causal Model was first called that by Holland (1986) for a series of articles developing the (a) general use of potential outcomes to define causal effects, (b) formalization of the assignment mechanism with possible dependence on the potential outcomes, and (c) extension to Bayesian posterior predictive inference, as well as the provision of what is now standard terminology (Rubin, 1974, 1975, 1977, 1978, 1980); a text from this perspective is Imbens and Rubin (2015); an entry length summary is Imbens and Rubin (2008); and a chapter length overview is Rubin (2008a), which includes more details on the Bayesian approach than given here. The use of potential outcomes in the context of what is now known as completely randomized experiments is due to Neyman (1923).

Fisherian inference is due to Fisher (1925) and the famous text by Fisher (1935) on experimental design, and is more fully developed by important followers such as Kempthorne (1952) and Wilk and Kempthorne (1956). Cox (1958) is a wonderfully clear exposition of many critical ideas in experimental design. Other texts on classical experimental design include, among others, Cochran and Cox (1957) and Box, Hunter, and Hunter (2005). Neymanian inference is due to Neyman (1923, 1935), of Neyman–Pearson fame. Recent texts applying classical frequentist approaches to observational studies include Rosenbaum (2002, 2010).

The need to design observational studies in analogy with the design of randomized experiments, appears to be a surprisingly recent emphasis; see Rubin (2007, 2008b), which is reinforced by D’Agostino and D’Agostino (2007), among others. It is instructive to compare this emphasis with earlier sage works, such as Cochran (1965, 1968). The terminology of propensity scores is due to Rosenbaum and Rubin (1983a), and Rubin (1997) provides a revealing example of its application; also see the book by Rubin (2006a) on matched sampling, and for a recent example see Fairlie, Zell, and Schrag (2013); for sensitivity analyses, see Rosenbaum and Rubin (1983b) and Rosenbaum (2010).

The terminology of principal stratification started with Frangakis and Rubin (2002); it is a method that generalizes far older ideas from economics called “instrumental variables” – see Angrist, Imbens, and Rubin (1996) for some history; also see Imbens and Rubin (1997), and Hirano, Imbens, Rubin, and Zhou (2000), as well as Somer and Zeger (1991), which was unaware of the economic history. Now principal stratification has many applications (e.g., Frangakis et al., 2004; Long, Little, & Lin, 2010; Baker & Lindeman, 1994; Mealli, Imbens, Ferro, & Biggeri, 2004; Rubin & Zell, 2010; and Rubin, Wang, Yin, & Zell, 2010).

For posterior predictive p -values, see Rubin (1984). For work on algorithms, see Dempster, Laird, and Rubin (1977), Gelman, Carlin, Stern, & Rubin (2004), and many other recent texts on Markov chain Monte Carlo (MCMC) methods.

For missing data see Little and Rubin (1987, 2002); Rubin (1987, 2004); Dorey, Little, and Schenker (1993); Little and Yau (1996); Wang, Little, Nan, and Harlow (2011); and Little et al. (2012). For partially defined outcomes, see Rubin (2006b).

Note

- 1 Throughout, matrices, or special cases such as row vectors or column vectors are set in boldface. When a symbol refers unambiguously to a scalar quantity, such as W_i , which can only be 0 or 1, it is not considered a special case of a matrix.

References

- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, 91, 444–455.
- Baker, S. G., & Lindeman, K. S. (1994). The paired availability design: A proposal for evaluating epidural analgesia during labor. *Statistics in Medicine*, 13, 2269–2278.
- Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for experiments: Design, innovation, and discovery* (2nd ed.). Hoboken, NJ: Wiley.
- Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A*, 128, 134–155.
- Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 295–313.
- Cochran W. G., & Cox, G. M. (1957). *Experimental designs*. New York, NY: Wiley.
- Cox, D. R. (1958). *Planning of experiments*. New York, NY: Wiley.
- D’Agostino, R. B., Jr., & D’Agostino, R. B., Sr. (2007). Estimating treatment effects using observational data. *JAMA*, 297, 314–316.
- Dempster, A. P., Laird, P., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data using the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Dorey, F. J., Little, R. J. A., & Schenker, N. (1993). Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine*, 12, 1589–1603.
- Fairlie, T., Zell, E. R., & Schrag, S. (2013). Effectiveness of intrapartum antibiotic prophylaxis from prevention of early-onset group B streptococcal disease. *Obstetrics and Gynecology*, 121, 570–577.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.
- Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Mahboobeh, S., Vlahov, D., & Strathdee, S. A. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *Journal of the American Statistical Association*, 99, 239–249.
- Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58, 21–29.
- Gelman, A., Carlin, J., Stern, H., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). New York, NY: CRC Press.
- Hirano, K., Imbens, G., Rubin, D. B., & Zhou, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1, 69–88.

- Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960.
- Imbens, G., & Rubin, D.B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*, 25, 305–327.
- Imbens, G., & Rubin, D.B. (2008). Rubin causal model. In S.M. Durlauf & C.E. Blume (Eds.), *The new Palgrave dictionary of economics* (2nd ed., Vol. 7, pp. 255–262). New York, NY: Palgrave MacMillan.
- Imbens, G., & Rubin, D.B. (2015). *Causal inference in statistics, and in the social and biomedical sciences*. New York, NY: Cambridge University Press.
- Kempthorne, O. (1952). *The design and analysis of experiments*. New York, NY: Wiley.
- Little, R. J., D’Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., . . . Stern, H. (2012). Special report: The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367, 1355–1360.
- Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.
- Little, R.J.A., & Rubin, D.B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.
- Little, R.J.A., & Yau, L. (1996). Intent-to-treat analysis in longitudinal studies with drop-outs. *Biometrics*, 52, 1324–1333.
- Long, Q., Little, R.J.A., & Lin, X. (2010). Estimating causal effects in trials involving multitreatment arms subject to non-compliance: A Bayesian framework. *Journal of the Royal Statistical Society, Applied Statistics*, 59, 513–531.
- Mealli, F., Imbens, G. W., Ferro, S., & Biggeri, A. (2004). Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics*, 5, 207–222.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. Translated in *Statistical Science*, 5, 465–480 (1990).
- Neyman, J. (1935). Statistical problems in agricultural experiments. Supplement to *Journal of the Royal Statistical Society, Series B*, 2, 107–108.
- Rosenbaum, P.R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer-Verlag.
- Rosenbaum, P.R. (2010). *Design of observational studies*. New York, NY: Springer-Verlag.
- Rosenbaum, P.R., & Rubin, D.B. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Rosenbaum, P.R., & Rubin, D.B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B*, 70, 41–55.
- Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66, 688–701.
- Rubin, D.B. (1975). Bayesian inference for causality: The importance of randomization. *Proceedings of the Social Statistics Section of the American Statistical Association*, 233–239.
- Rubin, D.B. (1977). Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2, 1–26.
- Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 34–58.
- Rubin, D.B. (1980). Discussion of “Randomization Analysis of Experimental Data in the Fisher Randomization Test” by Basu. *Journal of the American Statistical Association*, 75(371), 591–593.
- Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151–1172.
- Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.
- Rubin, D.B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127, 757–763.
- Rubin, D.B. (2004). *Multiple imputation for nonresponse in surveys (Wiley Classics Library)*. New York, NY: Wiley.
- Rubin, D.B. (2006a). *Matched sampling for causal effects*. New York, NY: Cambridge University Press.
- Rubin, D.B. (2006b). Causal inference through potential outcomes and principal stratification: Applications to studies with ‘censoring’ due to death. *Statistical Science*, 21, 299–321.
- Rubin, D.B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20–30.
- Rubin, D.B. (2008a). Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. In C. R. Rao, J. P. Miller, & D. C. Rao (Eds.), *Handbook of statistics: Epidemiology and medical statistics* (pp. 28–63). Amsterdam: Elsevier.

- Rubin, D. B. (2008b). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, 2, 808–840.
- Rubin, D. B., Wang, X., Yin, L., & Zell, E. R. (2010). Bayesian causal inference: Approaches to estimating the effect of treating hospital type on cancer survival in Sweden using principal stratification. In A. O'Hagan & M. West (Eds.), *The Oxford handbook of applied Bayesian analysis* (pp. 679–718). Oxford: Oxford University Press.
- Rubin, D. B., & Zell, E. R. (2010). Dealing with noncompliance and missing outcomes in a randomized trial using Bayesian technology: Prevention of perinatal sepsis clinical trial, Soweto, South Africa. *Statistical Methodology*, 7, 338–350.
- Somer, A., & Zeger, S. L. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine*, 10, 45–52.
- Wang, C., Little, R. J., Nan, B., & Harlow, S. (2011). A hot-deck multiple imputation procedure for gaps in longitudinal recurrent event histories. *Biometrics*, 67, 1573–1583.
- Wilk, M. B., & Kempthorne, O. (1956). Some aspects of the analysis of factorial experiments in a completely randomized design. *Annals of Mathematical Statistics*, 27, 950–985.

15

QUASI-EXPERIMENTAL DESIGN IN NURSING RESEARCH

Patricia Eckardt and David Rindskopf

There are three great lies. The check is in the mail. I'm from the government and I'm here to help you. I am going to randomly assign people to conditions.

—*Anonymous, as cited in Shadish, 2002, p. 4*

True experiments are characterized by randomization, manipulation, and control (comparison), and are used to determine causal effects of conditions being manipulated (Polit & Beck, 2011). In contrast to experiments, quasi-experiments lack random assignment to treatment conditions (Campbell & Stanley, 1963). Quasi-experimental designs are the result of unavoidable constraints arising from ethical, financial, or logistical limitations (Rosenbaum, 2002). Nursing science has many questions that are not appropriately answered using experiments. For example, to test the effect of delay in seeking treatment for acute heart failure (Jurgens, 2006), it would be impractical to randomly assign subjects to delay or nondelay in treatment seeking, and unethical if it were practical to do so.

Practical and ethical dilemmas can be avoided with quasi-experimental design, but causal inference is more challenging. Causal inference is the estimation of a cause and effect relationship between variables of interest (Cochran, 1965; Little & Rubin, 2000; Rosenbaum, 2002). Consideration of the challenges in the use of quasi-experiments for causal inference are important because causal questions are often posed in nursing science, and accurate answers are critical to optimizing patient outcomes when findings are applied in clinical practice.

This chapter provides an overview of quasi-experimental design and issues relevant to use of quasi-experimentation for making meaningful inferences about cause and effect in nursing science. The first section is an introduction to quasi-experimental (nonrandomized) design; subsequent sections provide an overview of the approaches of Donald Campbell and Donald Rubin to consideration of causal questions when they cannot be answered using randomized experiments (West & Thoemmes, 2010). The Campbellian perspective emphasizes selection of design features to rule out alternative explanations (Campbell, 1957; Campbell & Stanley, 1963). Rubin's perspective is based on the notion of potential outcomes (a counterfactual; the outcome that would have occurred for an individual under a different treatment, had it been assigned), and can be used to motivate approaches to modeling treatment selection mechanism when randomization is not possible (Coffman, this volume; Rubin, 1974, 1986, 2007; Rubin &

Zell, this volume). The final section summarizes four strategies to enhance causal inference in nonexperimental situations: matching, propensity scores, regression discontinuity, and the difference in differences approach.

Quasi-experiments

Terminology and background

Quasi-experimental designs in the nursing, social science, and educational literature are known as observational studies in the econometric, epidemiological, and bio-statistical sciences (Guo & Fraser, 2010). The term “observational studies” also applies to research that does not involve manipulation of treatment factors; however, some observational studies constitute a broader class of studies than quasi-experimental studies. Though the terminology surrounding nonrandomized experiments is not standardized, it is congruous; therefore, both sets of terminology will be used within this chapter.

Interest in causal inference from designed observational studies is longstanding in nursing (e.g., “Hand disinfection,” 1909). In statistics, Fisher (1925) explicated the rationale and analyses that link causal inference to the physical randomization of units to conditions in true experiments – with resultant statistical effect estimates that are unbiased, consistent, and efficient (Shadish, 2002). Subsequently, improved methodologies across the quantitative disciplines have provided new insights about justifiable causal inferences from data obtained in quasi-experiments (Heckmann, & Robb, 1989; Little & Rubin, 2000; Pearl, 2010; Shadish, Cook, & Campbell, 2002; Rindskopf, 1986; Rosenbaum, 2002; Winship & Morgan, 1999).

Causal inference from quasi-experiments must meet the requirements of all causal relationships that: cause precedes effect (a hypothesized cause cannot occur at the same time or after an observed effect); there is covariance between cause and effect (a relationship between cause and effect is present); and alternate explanations for the causal effect are implausible (other possible reasons for the observed relationship are not probable; Shadish et al., 2002). To examine the requirements of causal inference within quasi-experimental research, Campbell (1957) published a logical typology of quasi-experimental designs, and identified intrinsic strengths and limitations related to internal and external validity for each type of design. This literature, now often referred to as Campbellian (Shadish et al., 2002), is a common framework for teaching nonexperimental research in nursing research texts (e.g., Melnyk & Fineout-Overholt, 2010; Polit & Beck, 2011). The potential outcomes perspective with assignment mechanism (also known as Rubin’s causal model; Rubin, 2008; Rubin & Zell, this volume) provides a statistical framework for estimation of causal effects within experimental and observational studies.

Despite differences in approach, similarities between the Campbellian and Rubin perspectives regarding the importance of well-planned design are evident (West & Thoemmes, 2010). Commenting on the importance of design, Rosenbaum (2010) prefaced his book *Design of Observational Studies* with: “The quality and strength of evidence provided by an observational study is determined largely by its design. Excellent methods of analysis will not salvage a poorly designed study.” Shadish, Cook, and Campbell (2002) prefaced their text *Experimental and Quasi-experimental Designs for Generalized Causal Inference* with: “Recent statistical developments concerning causal inference in observational data have advanced understanding enormously. We want this book to complement such statistical work by emphasizing that, in the interplay between design and statistics, design rules!” In this chapter, we emphasize the design factors that support valid causal inference in quasi-experimental designs.

Campbellian quasi-experimental design

Campbell often referred to quasi-experiments as “queasy” experiments, noting the inherent messiness and greater susceptibility to inferential errors that quasi-experiments possess when contrasted to randomized experiments (Trochim, 1998). This is why consideration of potential threats to validity is so important in quasi-experiments.

Campbell (1957) distinguished randomized experimental, quasi-experimental, and nonexperimental (pre-experimental) research designs, and identified threats to validity (reasons why valid causal inference may not be plausible) to allow stronger design and analysis plans prior to collecting data (Shadish et al., 2002). Internal validity, external validity, and statistical conclusion validity were subsequently defined (Campbell & Stanley, 1963; Cook & Campbell, 1979): *Internal validity* refers to the extent to which the observed outcome can be attributable to the experimental treatment (i.e., intervention or manipulation). *External validity* concerns generalizability of inferences about effects across populations, settings, treatment variations, and alternate approaches to measurement of outcome variables. *Statistical conclusion validity* is the accuracy of the inference made about the population relationship between treatment and outcome based on sample data. In the Campbellian tradition, construct validity is the generalization of inference from a research study’s units and sample to general higher-order concepts, which is dependent upon an accurate estimation of covariances among variables and requires an accurate estimate of the covariance relationships in that study through statistical conclusion validity.

Inferential threats in quasi-experiments

Threats to internal validity are listed in Table 15.1. Examination of threats to validity needs to be exhaustive in order to rule out plausible alternative explanations of observed effects by design. Systematic consideration of each potential threat is essential to minimizing or avoiding threats. These considerations include the identification and examination of plausible threats to internal validity (other events or circumstances that could affect the accuracy of the study); primacy of control of threats to validity by design (which of these possible events or circumstances the researcher can control); coherent pattern matching (similarities of characteristics of groups being compared); and the construction of complex predictions to test causal hypotheses that few other alternate hypotheses could explain (predictive hypotheses that are as intricate as the research question).

Permanence of treatment effects is a critical consideration in the design of quasi-experiments. The effect of taking an aspirin for recurrent headache would be expected to be temporary, so that onset of a new headache would require repeating the treatment. By comparison, the effect of instruction about learning to solve an algebraic equation would be expected to be (relatively) permanent. Campbellian methods are primarily designed for study of interventions with permanent effects. We will primarily deal with the case of treatments with permanent effects as they are very common in nursing research and because effects of treatments with temporary effects would need discussion in a chapter of their own.

Results from true experiments possess internal validity by virtue of random assignment to treatment conditions. Randomization removes systematic bias arising from measured and unmeasured covariates by balancing these factors across treatment groups in expectation (Rubin & Zell, this volume). Intuitively, this important property from statistical theory means that if the experiment could be repeated an infinite number of times, the measured and unmeasured covariates would even out, on average.

Table 15.1 Factors jeopardizing internal validity

<i>Threat</i>	<i>Definition</i>	<i>Examples/Comments</i>
History	Specific events occurring between the first and second measurements in addition to the experimental variables	Changes to an electronic medical record documentation process occur during a study comparing efficiency with charting
Maturation	Processes within the participants that occur as a function of passage of time (not specific to particular events) that could account for the observed outcome	Growing older; becoming hungrier, more tired
Testing	Effects of taking a test upon scores obtained in a subsequent testing	When the same posttest is given that was the pretest, sufficient time has passed for recall of knowledge that were not readily recalled during pretest
Instrumentation	Changes in calibration of a measurement tool or changes in the observers or scorers that may account for the obtained measurements	An instrument measuring mass is calibrated daily at start of BMI study, but daily recalibration is neglected by some researchers due to process knowledge gap
Statistical regression	Tendency of a series of measurement to gravitate toward the mean of a distribution if the preceding measure was an extreme representative of that distribution	Operating where groups have been selected on the basis of their extreme scores
Selection biases	The nonrandom selection of participants from a population for study samples	Resulting from differential selection of respondents for the comparison groups
Experimental mortality	Differential loss of respondents from the comparison groups	Changes in primary care providers during a multisite study, patients who get admitted to hospital during an out-patient study
Selection-maturation interaction	The differences in the growth and development between the nonrandomly selected members of the comparison groups	Can occur in multiple-group quasi-experiments

Note: Threats were originally listed in Campbell and Stanley (1963).

Because quasi-experiments do not involve randomization to treatment conditions, the plausibility of alternative explanations of causality must be given serious and complete consideration. The Campbellian typology classifies quasi-experimental designs by resistance to threats to internal validity. Basic classes of quasi-experimental designs are: (a) single group designs that lack control groups; (b) control group designs that lack a pretest measure; and (c) designs with both control groups and pretests. Complex designs that simultaneously minimize multiple threats can be created by combining many elements of the basic designs. Table 15.2 lists examples of quasi-experiments constructed using principles discussed in this chapter.

Table 15.2 Design variation exemplars

<i>Design Variation</i>	<i>Example</i>
One-group: Add a pretest	In a culturally appropriate nine-month intervention to improve health behaviors in Hispanic mother–child dyads, use of a pretest allowed estimation of change in consumption of sugar-sweetened beverages and physical activity (Bender, Nader, Kennedy, & Gahagan, 2013)
One-group: Add a contrasting outcome	When examining partner aggressiveness on alcohol problems and affective functioning, laboratory cardiovascular responses to stressors were added to explore hypothesized pattern matching (Lorber, Erlanger, & Slep, 2013)
Bracketing	To explore the effect of posttraumatic stress disorder on spirituality in war veterans, nonmilitary matched control groups and differing war-era veterans were compared (Currier, Drescher, & Harris, 2014)
Comparison groups: Add double pretests	In a smoking cessation evaluation, double pretests allow consideration of maturational differences when estimating treatment effect (Shiyko, Li, & Rindskopf, 2012)
Switching replications	In an investigation of the effect of cognitive behavioral stress management on antiviral markers and mood, researchers wait-listed the randomized controls and then later provided treatment (Carrico et al., 2005)
Reversed treatment	In a classic quasi-experimental analysis, distinct groups were identified with reverse effects from treatment observed with a redesign of clerical jobs after technological advances at a bank (Hackman, Pearce, & Wolfe, 1978)
Sequential cohorts	To compare the outcomes of extreme premature births after multiple major advancements in technology, researchers used cohorts of babies born twenty years apart. (Bode et al., 2009)
Matching	In a “natural experiment,” the effect of obesity on lipids and blood pressure was evaluated by comparing heavier and lighter monozygous twins (“matched” on genotype) (Hayman, Meininger, Coates, & Gallagher, 1995)
Stratification	Elderly users of emergency department patients in Dublin were stratified on disposition after emergency care (admitted or discharged home) to profile users of the emergency department (Fealey et al., 2012)
Regression discontinuity	Effects of shorter or longer treatment for persons with alcoholism was studied by assigning treatment length on the basis of alcoholism severity (Flam-Zalcman et al., 2013)
Difference in differences	Effects of health care reform in Massachusetts over time were estimated by comparing outcomes in nearby states (Van der Wees, Zaslavsky, & Ayanian, 2013)

Quasi-experimental designs

Single group designs with no control group

Single group designs lack control groups (Table 15.3). They can reduce (but not eliminate) the plausibility of alternative explanations for a treatment effect (Trochim, 1998), because an effect may be observed but the cause is not necessarily observed. Three strategies can be used. First,

Table 15.3 Single group quasi-experimental designs

Occasions ^a			Dependent Variables ^b			Design	Measurement-Treatment Sequence
Treatment	Pretest	Posttest	Primary	Contrasting ^c			
1	0	1	1	0		Posttest only	X O ₁
1	0	1	1	≥ 1		Posttest only with multiple outcome variables	X O _{1A} , O _{1B}
1	1	1	1	0		Pre-post	O _{-1A} X O _{1A}
1	≥ 1	1	1	0		Pre-post with multiple pretest occasions	
1	1	1	1	≥ 1		Pre-post with multiple outcome variables	O _{-1A} , X O _{1A} , O _{-1B} O _{1B}
1	1	1	1	0		Pre-post with treatment removed	O _{-1A} X O _{1A} O _{2A} O _{3A}
≥ 1	1	1	1	0		Pre-post with treatment repeated	O _{-1A} X O _{1A} O _{2A} X O _{3A}

Note: A = primary dependent variable; B = contrasting outcome variable; O = outcome; X = treatment. Measurement occasions shown as negative numbers are made prior to treatment.

^aCell entries are the number of occasions for each portion of the study. ^bCell entries are the number of dependent variables. ^cContrasting outcomes are outcomes that are not indicators of the outcome construct of interest.

one or more pretests can be added to examine ongoing maturational trends. Second, multiple dependent variables that are indicators of the primary outcome and some unrelated construct can be measured, and pattern matching logic used to determine whether an apparent effect occurred in the theoretically expected outcome but not the contrasting variable (Coryn & Hobsin, 2011). Third, treatment can be removed or repeated if no learning or carryover effects are expected (Martin & Pear, 2006); this approach is often used in single-subject studies.

Control group designs with no pretest

Designs with a control group provide a comparison sample for the treatment group. In view of the Rubin causal model (Rubin & Zell, this volume), and under randomization to treatment and control conditions, outcomes in the control group provide information about potential outcomes that would occur had treatment not been received (i.e., the counterfactual situation) and allows unbiased estimation of population average treatment effect. Control groups play a similar role in quasi-experimental design, except inference arising from comparison of outcomes in treatment and control groups is subject to bias because the treatment assignment mechanism is not known when randomization is not used.

The posttest-only with nonequivalent groups is the basic design in this group (Table 15.4). Lack of a pretest measure limits availability of information about selection bias arising from how subjects were systematically assigned (or self-selected) to treatment levels. Design improvements include use of pretest in an independent sample, which provides useful pretreatment information when pretest measurements may be reactive, and when it is possible to randomly sample from

Table 15.4 Nonrandomized designs with a control group and no pretest or unconventional pretest design

<i>Pretest</i>					
<i>Sample</i>	<i>Variable</i>	<i>Design</i>	<i>Measurement-Treatment Sequence</i>		
None	None	Posttest only with nonequivalent groups	NR NR	X	O ₁ O ₁
Independent	Study outcome variable	Posttest with independent sample pretest ^a	NR NR	O ₋₁ X O ₋₁	O ₁ O ₁
Study participants	Proxy	Posttest with proxy pretest	NR NR	O _{-1A} X O _{-1A}	O _{1B} O _{1B}

Note: NR = not randomized to treatment or control group; A = primary dependent variable; B = proxy variable; O = outcome; X = treatment whereas absence of the X indicator denotes control or contrasting treatment. Measurement occasions shown as negative numbers are made prior to treatment.

^aItalics and the dashed vertical separator indicate that the pretest sample is different from the study sample that received the treatment and provided outcome data.

the population that was treated. Use of proxy pretests that are related conceptually to the outcome, and correlated to intervention, also improve inference, compared to the posttest only with nonequivalent groups design. Proxy pretests can provide information on selection bias differences between groups and also attrition bias between and within groups. Use of multiple-control groups that differ from each other, as much as they differ from the treatment group, can reveal hidden bias that may be the actual cause of observed differences on an outcome (Rosenbaum, 2002).

Bracketing involves preselection of two control groups, one of which is expected to score higher on the outcome than the treatment group and the second of which is expected to perform lower on the outcome variable than the treatment group. In a bracketed design, causal inference about the treatment effect is strengthened if the treatment group scores midway between both of the bracketed control groups (Shadish et al., 2002).

Designs with both control groups and pretests

Carefully selected control groups provide means for assessing a hypothesized counterfactual for a treatment group; addition of a pretest measure of the posttest variable adds to the plausibility of causal inference because it provides an estimate of the starting point for all participants, and exposes baseline differences in the groups (Shadish et al., 2002). In these designs, groups are nonequivalent by definition, as they were not drawn from the same population and were not randomly assigned to a treatment level, resulting in an assumption of selection bias into treatment groups (Shadish et al., 2002). Use of a pretest allows investigation of the magnitude and direction of the selection bias that is assumed.

Some variations on the commonly used nonequivalent comparison group design with pretest and posttest are shown in Table 15.5. Use of a *double pretest* on two occasions prior to intervention allows detection of maturation differences between the groups. *Switching replications of treatment* involves providing treatment to the control group at a later time. This replication is not exact because time elapses and maintaining fidelity of treatment administration may be challenging.

Table 15.5 Nonrandomized designs with both control groups and pretests

Design	Measurement-Treatment Sequence					
Untreated control with pretest/posttest	NR	Treatment	O_{-1}	X	O_1	
	NR	Control	O_{-1}		O_1	
Untreated control with series pretest and single posttest	NR	Treatment	$O_{-j}, \dots, O_{-1}$	X	O_1	
	NR	Control	$O_{-j}, \dots, O_{-1}$		O_1	
Untreated control with pretest/posttest and switching replications ^a	NR	Treatment	O_{-1}	X	O_1	O_2
	NR	Control	O_{-1}	X	O_1	$X_2 \quad O_2$
Untreated control with pretest/posttest and reversed treatment/control	NR	Treatment	O_{-1}	X_+	O_1	
	NR	Control	O_{-1}	X_-	O_1	
Cohort control	NR	Treatment		X	O_1	
	NR	Control	O_{-1}			
Cohort control with pretest in each cohort	NR	Treatment	$O_{-j}, \dots, O_{-1}$	X	O_1	
	NR	Control	$O_{-j}, \dots, O_{-1}$			

Note: NR = not randomized to treatment or control group; O = outcome; X = treatment whereas absence of the X indicator denotes control or contrasting treatment. Measurement occasions shown as negative numbers are made prior to treatment; j = the first in a series of pretests prior to the time of intervention in the treatment group.

^aAdministering the experimental group's treatment to the control group at a later time.

However, this design may be a reasonable option when funding is limited or research resources are scarce.

In the Campbellian tradition, cohorts are intact groups experiencing events, including treatments, at successive times and with a *cohort design* then being a quasi-experimental design arising from comparison of outcomes in cohorts before and after a treatment was instituted. Cohort designs are useful if cohorts remain similar over time in composition and other variables that may affect the outcome. Administration of a pretest strengthens cohort designs because magnitude and direction of selection bias can be exposed. However, cohorts may change over time or a historical event may occur that affects the outcome variable distinct from the treatment variable being studied.

When pretests are used in conjunction with control groups, the pretest may reveal ceiling or floor effects if a test of the outcome is too easy or too difficult, which would result in inability to discriminate among people with high or low performance on the outcome of interest (Allen & Yen, 1979/2002; Glass & Hopkins, 1996). Ideally, these effects should be explored in a pilot study rather than at the time of the pretest in a full-scale study. Failure to find any difference on the pretest does not demonstrate equivalence of the groups, whereas existence of pretest differences is a signal for selection bias alone or through interaction with other threats. For example, when selection bias and instrumentation interact (e.g., control group pretest average scores are lower and ceiling effects are operative in the experimental group), a control group average increase at posttest is more readily detected, and the posttest difference between the two groups is artifactually similar.

Complex quasi-experimental designs incorporate variations in pretests and control groups to attempt to control multiple threats to causal inference and strengthen causal inference. Complex designs account for as many threats to validity as practically possible. There is no perfect design for answering a research question, however. Units (participants) to be contrasted, treatments, observations made on the units, and the settings in which the study takes place all need to be considered to construct the best design. Optimal design allows generalization to units, treatments,

variables, and settings not directly observed with proposed design (Shadish et al., 2002). During the planning stage, research designs need to be probed for limitations arising from threats to validity, and proposed design changes incorporated prior to any research being conducted. The purpose of this planning work is to decrease the identified potential threats and increase the plausibility of any causal inferences from estimates of effect. Even with the most thoughtful consideration to management of threats to inference by design, alternate explanations may be plausible and should always be considered when interpreting findings.

Rubin's potential outcomes-based causal model

A counterfactual is an outcome that would have occurred for an individual if they had in fact received the treatment they did not receive in a research study, and can be referred to as a “potential outcome” (Hill, Weiss, & Zhai, 2011; Winship & Morgan, 1999). Within the experimental design literature, the counterfactual account of causality has been explored for many years (Cochran & Cox, 1950; Fisher, 1935; Neyman, 1923/1990, Iwaskiewicz, & Kolodziejczyk, 1935). Rubin (1974, 1977, 1981, 1986, 1990, 2008) formalized the causal account in nonexperimental designs using a potential outcomes framework. Econometricians also applied the potential outcomes account of causality (Heckman & Robb, 1989; Imbens, 2010).

Rubin's potential outcomes framework involves conceptualization of the data from a quasi-experiment as “having arisen from a complex randomized experiment, where the rules used to assign the treatment conditions have been lost and must be reconstructed” (Rubin, 2008, p. 815). Two essential parts of the potential outcomes framework are: (a) defining causal effects; and (b) modeling treatment assignment mechanism. Implementing the approach may also involve specifying a full probability model for the factors used to predict treatment assignment (“the science”), which are treated as fixed in assignment-based approaches to predicting outcomes (Rubin, 2008, p. 815; Rubin & Zell, this volume).

Defining causal effects within the potential outcomes framework

Key points in the definition of causal effects are: (a) each study unit (participant) either receives the treatment or does not receive the treatment (is a control), and there is no effect on any unit's potential outcome due to the treatment assignment of another unit (the stable unit treatment value assumption [SUTVA]); (b) the potential outcome is what would have occurred under the treatment condition that was not experienced; and (c) the *causal effect for the unit* is the difference at some defined point in time between the observed and potential outcome (when the outcome varies continuously) (Rubin & Zell, this volume). Naturally, only actual outcomes are observed. In a randomized experiment, the observed outcomes of subjects in treatment and control conditions can be compared to estimate the *population average treatment effect* conditioned on the known, randomly assigned treatment value ($0 = \text{control}$; $1 = \text{treatment}$; Hill, Weiss, & Zhai, 2011). In quasi-experiments, treatment assignment mechanism is not known and is not random by design, so special considerations are needed.

Treatment assignment mechanisms in quasi-experiments

In true experiments, randomization to treatment assignment (experimental or control group) serves to “balance” the distribution of measured and unmeasured covariates that may affect the outcome of interest. In quasi-experiments, it is known for certain that the treatment assignment mechanism did not involve randomization, and it is even possible that treatment assignment may

be a function of potential outcomes (Rubin, 2008, p. 814). Often, however, the decision rules for treatment assignment (evidence base, personal preference), who made them (care provider, patient, patient family), and degree of fidelity to treatment assignment rules by all decision makers are also known. Availability of information about key covariates (other plausible factors influencing treatment assignment) is essential to specifying the treatment assignment mechanism. Conceptually, the goal is to use available decision rule information and the covariates to estimate the probability of assignment to treatment or control conditions. In research practice, information about treatment assignment mechanism is used to determine whether balance in treated and control units can be achieved by creating similar subclasses or matched pairs for the purpose of comparing outcomes under the treatment conditions.

Special design variations

Matching and stratifying

Matching pairs or matching subgroups (strata) using one or a small set of covariates creates balance in treated and control units aimed at allowing the matched units or subgroups to serve as counterfactual proxies (e.g., Samuels & Eckardt, 2013). Many matching methods are available; some match multiple controls to one treated subject and most allow approximate rather than exact matching (Rosenbaum & Rubin, 1984). Simple “one-to-one matching” pairs a control subject and treatment subject who are similar on pre-intervention covariates hypothesized to affect the outcome. By matching pairs on characteristics such as gender, age, blood pressure, illness history, and so on that might be related to outcome, it becomes more plausible that the members of a pair would have been the same on the outcome – except for the effect of treatment.

Undermatching occurs when some important covariates are not included in the matching process. More important is the problem of nonoverlapping of the groups on the matched variable of interest. If there is not sufficient overlap in covariates between these in treated and control units, then only local treatment effects in the range where there is sufficient overlap should be estimated (Gelman & Hill, 2006). If there is no overlap of covariates in treated and control units, no comparisons should be made as there is no approximation to a counterfactual with the observed data.

Stratification is a method of controlling for systematic differences, other than the treatment group assignment, between units that a researcher wants to make some inference about. Stratification involves grouping those units into subclasses based on observed characteristics, and then directly comparing only treated and control units who fall within the same subclass. An example of stratification would be grouping the patients in a study testing the effect of motivational interviewing on dietary adherence into substrata based on the patients’ self-reported locus of control prior to the intervention. Stratification can reduce 90% of bias on observed covariates if a minimum of five strata are used (Cochran, 1965, 1968). However, as the number of covariates increase that the researcher needs to stratify on to control for these differences, samples can become too small for inference within a stratum, and sometimes consist of only members from one treatment group.

Propensity scores

The propensity score is defined as the conditional probability of assignment to the treatment group given a set of observed pretreatment variables (Eckardt, 2010, 2012; Gelman & Hill, 2006; Rosenbaum & Rubin, 1983, 1984). Propensity scores reduce multiple covariates to a single

composite score (Shadish et al., 2002). Propensity scores control for differences in treatment and control groups and create balance because the relationship between observed covariates and treatment assignment is independent, conditioned on the propensity score (Imai & van Dyk, 2004).

Propensity scores are known when units are randomized to treatment assignment (Rubin & Zell, this volume). In observational studies, propensity score estimates are obtained by regressing treatment assignment on all covariates believed to influence assignment to treatment. In the case of two treatment assignments (experimental and control), a logistic regression approach is used; the approach can be generalized to ordinal and categorical treatment levels (Imai & van Dyk, 2004).

Regular assignment mechanisms are unconfounded, probabilistic, and individualistic (Rubin & Zell, this volume). Assignment mechanisms are ignorable if they are regular. If ignorability holds in an observational study, propensity scores can be used to obtain unbiased treatment effect estimates. Propensity scores can be used in a variety of ways (Austin & Laupacis, 2011): to (a) match units in treatment and comparison groups, (b) create strata, (c) serve as a weight when inverse probability weighting of treatment is used, and (d) adjust for covariates. Details on estimating and using propensity scores are available in many sources (Austin & Laupacis, 2011; Guo & Fraser, 2010; Rubin, 2008). Using these approaches to analyze data from an observational study does not insure estimates of causal effects that are the same as if a randomized experiment had been conducted. These approaches do, however, create approximations to a randomized design that help to decrease bias arising from lack of randomization (Rubin, 2007, 2008).

Regression discontinuity design

The regression discontinuity design (RDD) is an observational approach that takes advantage of *systematic assignment* of treatment to units to obtain an unbiased estimate of effect (Imbens & Lemieux, 2007; Rubin, 1977; Thistlethwaite & Campbell, 1960; Trochim, 1984). In the RDD, treatment assignment is based solely on whether a pre-intervention measure is above or below an established threshold or cut point on a covariate called the forcing variable (Schochet et al., 2010). The regression of outcome on treatment assignment based on value of the forcing variable is estimated; the gap in regression lines at the cut point is the estimate of (local) treatment effect (Gelman & Hill, 2006). Correct specification of the functional form for the model linking the forcing variable to the outcome and selection of an appropriate range of the forcing variable are critical (Imbens & Lemieux, 2007). RDDs provide critical evidence about treatment effects when RCTs are not possible due to ethical or feasibility considerations, and are consistent with many behavioral, educational, and psychosocial intervention approaches in real-life applications (Johnston & Dijkers, 2012).

The RDD can be adapted by allowing random assignment to treatment for those near the established threshold (Battistin & Rettore, 2002) or a smaller jump in the probability of assignment to treatment at the threshold (Imbens & Lemieux, 2007). For thorough presentations of the regression discontinuity approach, see Shadish, Cook, and Campbell (2002, Chapter 7) and Winship and Morgan (1999).

Difference in differences

The idea in a difference in differences (DID) design is to compare control and treatment groups on an outcome with measures repeated before and after intervention in the treatment group. Repeated cross-sectional data and panel data (repeated measurements of individuals) may be used (Imbens & Wooldridge, 2007). The DID design with panel data is analogous to the nonequivalent

control group pretest–posttest design in Campbellian literature (Angrist & Krueger, 2001). The logic is that change in mean scores in the control group reflects the same change the treatment group would have experienced had they not been exposed to the treatment. Assumptions are that treatment effect is constant or additive (constant with noise) across individuals and that unobserved covariates have the same distribution across time points and treatment groups. Comparison across time points controls for other possible observed and unobserved differences across the groups being compared (Gelman & Hill, 2006). For resources, see Angrist and Krueger (2001) and Imbens and Wooldridge (2007).

Summary

Many research situations in nursing science involve nonrandomization of assignment to treatment groups to answer important causal questions. Sound methodological design and thoughtful approaches to analysis of observational studies are needed to guide the translation of strong evidence into practice. The approaches of Campbell and Rubin provide insights into design, analysis of data, and assessment of conclusions of quasi-experimental studies in nursing science.

References

- Allen, M. J., & Yen, W. M. (1979/2002). *Introduction to measurement theory*. Prospect Heights, IL: Waveland.
- Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *National Bureau of Economic Research*, No. w8456. Retrieved from www.nber.org/papers/w8456
- Austin, P. C., & Laupacis, A. (2011). A tutorial on methods to estimating clinically and policy-meaningful measures of treatment effects in prospective observational studies: A review. *International Journal of Biostatistics*, 7(1), Article 6. doi:10.2202/1557-4679.1285
- Battistin, E., & Rettore, E. (2002). Testing for programme effects in a regression discontinuity design with imperfect compliance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165, 39–57.
- Bender, M. S., Nader, P. R., Kennedy, C., & Gahagan, S. (2013). A culturally appropriate intervention to improve health behaviors in Hispanic mother–child dyads. *Childhood Obesity*, 9, 157–163.
- Bode, M. M., D'Eugenio, D. B., Forsyth, N., Coleman, J., Gross, C. R., & Gross, S. J. (2009). Outcome of extreme prematurity: A prospective comparison of 2 regional cohorts born 20 years apart. *Pediatrics*, 124(3), 866–874. doi:10.1542/peds.2008-1669
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297–312.
- Campbell, D. T., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.
- Carrico, A. W., Antoni, M. H., Pereira, D. B., Fletcher, M. A., Klimas, N., Lechner, S. C., & Schneiderman, N. (2005). Cognitive behavioral stress management effects on mood, social support, and a marker of antiviral immunity are maintained up to 1 year in HIV-infected gay men. *International Journal of Behavioral Medicine*, 12, 218–226.
- Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society: Series A (General)*, 128, 234–266.
- Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 295–313.
- Cochran, W. G., & Cox, G. M. (1950). *Experimental design* (2nd ed.). New York, NY: Wiley.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis for field settings*. Chicago, IL: Rand McNally.
- Coryn, C. L. S., & Hobson, K. A. (2011). Using nonequivalent dependent variables to reduce internal validity threats in quasi-experiments: Rationale, history, and examples from practice. In S. Mathison (Ed.), *Really new directions in evaluation: Young evaluators' perspectives*. *New Directions for Evaluation*, 131, 31–39.

- Currier, J. M., Drescher, K. D., & Harris, J. I. (2014). Spiritual functioning among veterans seeking residential treatment for PTSD: A matched control group study. *Spirituality in Clinical Practice*, 1(1), 3–15.
- Eckardt, P. A. (2010). *The application of propensity score estimates in hierarchical linear models for causal inference*. (Doctoral dissertation). Retrieved from Proquest. (Accession 3426733)
- Eckardt, P. (2012). Propensity score estimates in multilevel models for causal inference. *Nursing Research*, 61, 213–223.
- Fealey, G. M., Treacy, M., Drennan, J., Naughton, C., Butler, M., & Lyons, I. (2012). A profile of older emergency department attendees: Findings from an Irish study. *Journal of Advanced Nursing*, 68, 1003–1013.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.
- Flam-Zalcman, R., Mann, R. E., Stoduto, G., Nochajski, T. H., Rush, B. R., Koski-Jännes, A., . . . Rehm, J. (2013). Evidence from regression-discontinuity analyses for beneficial effects of a criterion-based increase in alcohol treatment. *International Journal of Methods in Psychiatric Research*, 22, 59–70.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.
- Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston, MA: Allyn & Bacon.
- Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage.
- Hackman, J. R., Pearce, J. L., & Wolfe, J. C. (1978). Effects of changes in job characteristics on work attitudes and behaviors: A naturally occurring quasi-experiment. *Organizational Behavior and Human Performance*, 21, 289–304.
- Hand Disinfection. (1909). *British Journal of Nursing*, 43, 251–252.
- Hayman, L. L., Meininger, J. C., Coates, P. M., & Gallagher, P. R. (1995). Nongenetic influences of obesity on risk factors for cardiovascular disease during two phases of development. *Nursing Research*, 44, 277–283.
- Heckman, J. J., & Robb, R. (1989). The value of longitudinal data for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In D. Kasprzyk, G. Duncan, G. Kalton, & M. P. Singh (Eds.), *Panel surveys* (pp. 512–538). New York, NY: Wiley.
- Hill, J., Weiss, C., & Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, 46, 477–513.
- Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99, 854–866.
- Imbens, G. W. (2010). An economist's perspective on Shadish (2002) and West and Thoemmes (2010). *Psychological Methods*, 15, 47–55.
- Imbens, G., & Lemieux, T. (2007). *Regression discontinuity designs: A guide to practice*. Working Paper 13039. National Bureau of Economic Research. Retrieved from www.nber.org/papers/w13039.pdf
- Imbens, G. W., & Wooldridge, J. M. (2007, July–August). What's new in econometrics? Difference-in-differences estimation. National Bureau of Economic Research, Lecture Notes. Retrieved from www.nber.org/minicourse3.html
- Johnston, M. V., & Dijkers, M. P. (2012). Toward improved evidence standards and methods for rehabilitation: Recommendations and challenges. *Archives of Physical Medicine and Rehabilitation*, 93(8 Suppl), S185–S199.
- Jurgens, C. Y. (2006). Somatic awareness, uncertainty, and delay in care-seeking in acute heart failure. *Research in Nursing & Health*, 29, 74–86.
- Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health*, 21, 121–145.
- Lorber, M. F., Erlanger, A. C., & Slep, A. M. S. (2013). Biological sensitivity to context in couples: Why partner aggression hurts some more than others. *Journal of Consulting and Clinical Psychology*, 81, 166–176.
- Martin, G., & Pear, J. (2006). *Behavior modification: What it is and how to do it* (8th ed.). Upper Saddle River, NJ: Prentice Hall.
- Melnik, B. M., & Fineout-Overholt, E. (2010). *Evidence-based practice in nursing & healthcare: A guide to best practice* (2nd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.
- Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (D. M. Dabrowska & T. P. Speed, Trans.). *Statistical Science*, 5, 465–480. (Original work published 1923)
- Neyman, J., Iwaskiewicz, K., & Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation (with discussion). *Supplement to the Journal of the Royal Statistical Society*, 2, 107–180.
- Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology*, 40, 75–149.

- Polit, D.F., & Beck, C. T. (2011). *Nursing research: Generating and assessing evidence for nursing practice* (9th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.
- Rindskopf, D. (1986). New developments in selection modeling for quasi-experimentation. *New Directions for Program Evaluation*, 31, 79–89.
- Rosenbaum, P.R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.
- Rosenbaum, P.R. (2010). *Design of observational studies*. New York, NY: Springer.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Rosenbaum, P.R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516–524.
- Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.
- Rubin, D.B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral Statistics*, 2, 1–26.
- Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational and Behavioral Statistics*, 6, 377–401.
- Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers? *Journal of the American Statistical Association*, 81, 961–962.
- Rubin, D.B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5, 472–480.
- Rubin, D. B. (2007). The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20–36.
- Rubin, D.B. (2008). For objective causal inference design trumps analysis. *Annals of Applied Statistics*, 2, 808–840.
- Samuels, J. G., & Eckardt, P. (2013). The impact of assessment and reassessment documentation on the trajectory of postoperative pain severity: A pilot study. *Pain Management Nursing*. Advance online publication. doi:10.1016/j.pmn.2013.07.007
- Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). Standards for regression discontinuity designs. Version 1.0 (Pilot). Retrieved from Institute for Education Sciences, What Works Clearinghouse at http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_rd.pdf
- Shadish, W.R. (2002). Revisiting field experimentation: Field notes for the future. *Psychological Methods*, 7, 3–18.
- Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Shiyko, M. P., Li, Y., & Rindskopf, D. (2012). Poisson growth mixture modeling of intensive longitudinal data: An application to smoking cessation behavior. *Structural Equation Modeling*, 19, 65–85.
- Thistlethwaite, D.L., & Campbell, D.T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51, 309–317.
- Trochim, W.M.K. (1984). *Research design for program evaluation: The regression-discontinuity approach (Contemporary evaluation research)*. Beverly Hills, CA: Sage.
- Trochim, W.M.K. (1998). Donald T. Campbell and research design. *American Journal of Evaluation*, 19, 407–409.
- Van der Wees, P. J., Zaslavsky, A. M., & Ayanian, J. Z. (2013). Improvements in health status after Massachusetts health care reform. *Milbank Quarterly*, 91, 663–689.
- West, S.G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*, 15, 18–37.
- Winship, C., & Morgan, S.L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, 25, 659–706.

16

THIRD VARIABLES

Scientific meanings and modeling in non-randomized studies

Donna L. Coffman

The relationship between two key variables can become considerably more complicated when a third variable is introduced. Third variables can act as moderators, mediators, confounders, and instrumental variables. Their scientific meaning determines their role and how they are modeled. The goal of this chapter is to scientifically distinguish among the various roles of third variables, particularly when attempting to estimate causal effects from data obtained in observational or quasi-experimental studies.

The chapter is organized as follows. First, I define terms and notation. I introduce the potential outcomes framework for causal inference. Then I discuss moderators, confounders, mediators, and instrumental variables in turn. Finally, I make some concluding remarks about how researchers should consider these variables in their research.

Definitions and notation

Consider an exposure or treatment variable, denoted T , and an outcome, denoted Y . Notation for a third variable differs depending on the role that it plays. These roles include moderator, denoted Z ; mediator, denoted M ; confounder, denoted X ; and instrumental variable, denoted IV . Moderators are variables that alter the effect of the exposure on the outcome; that is, the effect of the exposure on the outcome varies across levels of the moderator. Mediators are variables that lie in the causal pathway between exposure and outcome; the mediator transmits the effect of the exposure to the outcome. Confounders are variables that have an effect on both the exposure and outcome and, when not properly taken into account, bias the estimate of the effect of the exposure on the outcome. Instrumental variables strongly influence the exposure but do not have an effect on the outcome. I will discuss each in turn and then discuss how the same construct could be a mediator in one context and a confounder in another context, depending on the research question.

Throughout the chapter, I will use the terms *exposure*, *treatment*, and *intervention* interchangeably to refer to the variable that is a hypothesized cause. Also, unless otherwise noted, all figures are time-ordered from left to right. The phrase *to account for confounding* refers to either including the confounder as a covariate in a regression equation or including it in a propensity score model. Finally, the word *covariate* is used broadly to refer to any of the different types of third variables.

Potential outcomes framework

One important advantage of the potential outcomes framework for causal inference (Rubin, 1974; Rubin & Zell, this volume) is that it allows clear definition of causal effects. The potential outcomes framework is described in detail elsewhere (e.g., Holland, 1986; Little & Rubin, 2000; Rubin, 2005; Schafer & Kang, 2008; Sobel, 1995; Winship & Morgan, 1999). Briefly, each individual has a potential outcome under each treatment condition. The causal effect is defined as a contrast between potential outcomes for different treatment conditions for each individual. For example, consider a binary treatment indicator, T_i , where $T_i = 1$ denotes the intervention condition and $T_i = 0$ denotes the control condition for participant i , $i = 1, \dots, N$. In this case, there are two potential outcomes for each individual: the potential outcome if the individual receives the intervention, denoted $Y_i(1)$, and the potential outcome if the individual is in the control condition, denoted $Y_i(0)$. Because an individual can be observed in only one exposure condition, only one of these potential outcomes is observed; the other is missing (and sometimes referred to as a *counterfactual outcome*). Because of this, the individual causal effect, defined as the difference between the potential outcomes, $Y_i(1) - Y_i(0)$, cannot be estimated without further assumptions. When such assumptions (described below) are made, the average causal effect (ACE) across all individuals can be estimated. The ACE is defined as the expected difference between the potential outcomes in the population, $E[Y_i(1) - Y_i(0)]$. However, the ACE alone does not provide information about mechanisms or mediating pathways.

Moderators

As depicted in Figure 16.1, a moderator does not affect either the exposure or the outcome; rather, a moderator has an effect on the effect of the exposure on the outcome. In other words, the effect of the exposure on the outcome differs depending on the level of the moderator. Moderators may be categorical (such as gender), ordinal (such as grade in school), or continuous (such as age).

In linear models, moderation of a treatment effect is modeled by constructing a predictor variable that is the product of the treatment value and the hypothesized moderator. For example, if it is hypothesized that gender moderates the effect of an intervention, then the values of the gender indicator variable and the intervention indicator variable are multiplied together and included in the model as a predictor of the outcome. If the regression coefficient of this product term is statistically significant, then the moderation effect is supported. Interpretation of the regression coefficient of the product term depends on the type of coding (e.g., effect, dummy) that the analyst used (e.g., Cohen, Cohen, West, & Aiken, 2003).

Moderation becomes more complex in non-linear models. For example, for binary outcomes, moderation may be present when estimating the risk ratio but not when estimating the risk difference for the same data (Cohen et al., 2003). Not only does the interpretation of the regression coefficient for the product term vary depending on the scale of the outcome, but the presence of moderation may also depend on the scale of the outcome (VanderWeele, 2009b).

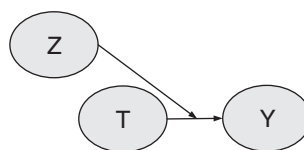


Figure 16.1 A moderator effect.

The role of time in moderation

The influential Baron and Kenny (1986) article distinguishing moderators and mediators did not impose a temporal precedence criterion. Others subsequently specified that moderators precede that which they moderate, whereas mediators follow that which they mediate (Kraemer, Stice, Kazdin, Offord, & Kupfer, 2001; Kraemer, Wilson, Fairburn, & Agras, 2002; Kraemer, Kiernan, Essex, & Kupfer, 2008). However, this temporal precedence restriction does not prohibit a particular construct measured at baseline as a moderator from being measured post-intervention as a mediator of the effect of the intervention on the outcome. In other words, the same construct measured on different occasions can play different roles.

Time-invariant moderators are variables such as race that do not change over time. Moderators can also vary across time (Almirall, Ten Have, & Murphy, 2010; Almirall, McCaffrey, Ramchand, & Murphy, 2013). Examples of questions involving time-varying moderators are “How does suicidal ideation history moderate the effect of early vs. later depression treatment on future depression severity?” and “How does time-varying aggressiveness moderate the effect of early vs. later grade retention on academic achievement?”

Consider Figure 16.2, in which the exposure and the moderator may vary over time. The subscripts denote baseline, time 1 and time 2. Figure 16.2, Panel A isolates the effect of the exposure at time 1 on the outcome. Note that only Z_0 moderates the effect of T_1 on Y because Z_1 occurs after T_1 and therefore cannot moderate the effect of T_1 on Y . T_2 is typically set to “off,” as represented by 0, but it could be set to any of its possible values. In any case, it is held constant when examining only the effect of initial treatment. Figure 16.2, Panel B illustrates the cumulative effect of exposure on the outcome. The cumulative effect is represented by the arrow from T_1 to the effect of T_2 on Y . That is, the initial treatment moderates the effect of T_2 on Y . In addition, Z_1 may also moderate the effect of T_2 on Y because it occurs prior to T_2 . Figure 16.2 is meant to be conceptual, in that there may be other effects that are not illustrated. For example, an arrow from Z_0 to Z_1 could be added.

Figure 16.2, Panel B also illustrates an important point regarding the differences between two concepts referred to as treatment interaction effects and effect modification in the epidemiology literature (e.g., VanderWeele, 2009b). Treatment interaction effects refer to the cumulative effect illustrated by the arrow from T_1 to the effect of T_2 on Y . In other words, treatment interactions are between two variables, both of which are hypothesized causes of the outcome. Effect

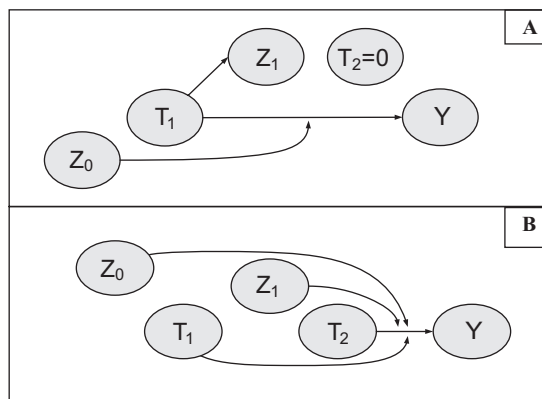


Figure 16.2 Time-varying moderator effects.

modification refers to the effects of Z_0 and Z_1 on the effect of T_2 on Y . In other words, product terms representing effect modification are between a hypothesized cause and a moderator that is not hypothesized to have a causal effect on the outcome. In both cases, product terms are included in the model but the scientific meaning is slightly different. In the behavioral sciences, this distinction is not typically made: all moderator effects or effects involving product terms are generally referred to as interaction effects.

Causal inference in presence of moderation

Appropriate modeling approaches for causal inference in the presence of moderation depend on whether the moderator is time invariant or time varying. Marginal structural models (MSMs; Robins, Hernan, & Brumback, 2000) can be used for time-invariant but not time-varying moderators. Coffman, Caldwell, and Smith (2012), for example, used this approach to model the moderating effect of baseline risk status on the effect of leisure boredom on initiation of cigarette smoking. Structural mean models are used in the presence of time-varying moderation. For a description of structural mean models and an application to treatment for depression, see Almirall et al. (2013) and Almirall et al. (2010).

Confounders

In the time-invariant situation, a confounder is a variable that has an effect on the outcome and is associated with the exposure; in the time-varying situation, a confounder is a variable that has an effect on the outcome as well as an effect on (not just an association with) the exposure (Shadish, Cook, & Campbell, 2002; Greenland & Morgenstern, 2001). As in the latter case, I will consider a confounder to be a variable that affects both the exposure and the outcome.

Figure 16.3 shows a confounder, X , of T and Y in the case where there is no causal effect of T on Y . If the confounder is not taken into account when estimating the effect of exposure or treatment on the outcome, a causal effect will appear, incorrectly, to be present between the exposure and outcome, even though the exposure does not actually have an effect on the outcome (depicted by the lack of an arrow between T and Y in Figure 16.3). A common example of confounding is the positive association between the number of churches in a city and the crime rate. Of course, the number of churches does not cause an increase in crime; rather, the two are related because an increase in population density causes an increase in both churches and crime.

The terms confounding and selection bias are often used interchangeably in the behavioral sciences but they are distinguished from each other in the epidemiology literature (Hernan, Hernandez-Diaz, & Robins, 2004). Figure 16.4 shows that selection bias results when a treatment exposure and outcome both affect a common outcome, denoted E ; as described above and shown in Figure 16.3, confounding arises when a covariate X affects both treatment and outcome.

Consider an example in which the exposure T is a genetic marker, the outcome Y is smoking, and both cause lung cancer, E . The question of interest is whether the genetic marker causes

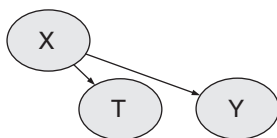


Figure 16.3 Confounding effects.

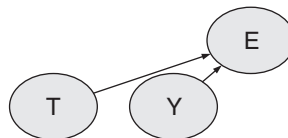


Figure 16.4 Selection bias.

smoking. Conditioning on the common effect (E) of the exposure (T) and outcome (Y) results in selection bias (Rosenbaum, 1984). For example, conditioning on having lung cancer (E), there may appear to be a causal effect between the genetic marker (T) and smoking (Y), even if no such effect exists. This would happen if among those who have lung cancer, non-smokers are much more likely than smokers to have the genetic marker. In fact, if these were the only two causes of lung cancer, non-smokers who have lung cancer must have the genetic marker but smokers who have lung cancer may or may not have the genetic marker.

Time-varying confounding

As in the case of moderators, temporal issues arise when considering confounders. Time-varying confounding arises when assessing the effects of a time-varying treatment; a time-varying confounder is a confounder of the effect of current treatment on the outcome when the confounder itself has been influenced by the past treatment history. Failing to account for a time-varying confounder will bias estimates of the effect of current treatment on a subsequent outcome. However, conditioning on a time-varying confounder will result in biased estimates of the cumulative effect of the treatment on the outcome. MSMs with inverse probability weighting have been proposed for estimating time-varying treatment effects in the presence of time-varying confounding (e.g., Barber, Murphy, & Verbitsky, 2004; Bray, Almirall, Zimmerman, Lynam, & Murphy, 2006; Robins et al., 2000). These models appropriately account for time-varying confounders without conditioning on them.

Mediators

Mediation is illustrated in Figure 16.5. Mediation analysis, by definition, involves questions about causal effects and mechanisms. The exposure T has an effect on the mediator M , which in turn has an effect on the outcome Y . This effect is referred to as the indirect or mediated effect. In addition, T may have an effect on Y that does not go through M , which is referred to as the direct effect (MacKinnon, 2008).

Judd and Kenny (1981) stressed that researchers should consider mediation analysis to be a correlational analysis due to measurement error, reciprocal causation, and the possibility of omitting important variables. In regression analysis the omitted variable problem results from confounders that are unmeasured or not accounted for. In particular, Judd and Kenny (1981, p. 607) stated, “If variables that affect the outcome and that also are causes of the included mediating variables are not controlled in the analysis, biased estimates of the mediation process will result, *even when a randomized experimental research design has been used.*” However, the later influential Baron and Kenny (1986) paper did not mention this issue at all. Only recently this issue has been revisited using the potential outcomes framework, primarily in the methodological literature (e.g., Coffman, 2011; Ten Have & Joffe, 2012; Rubin, 2004). Without further assumptions, direct

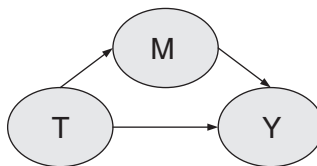


Figure 16.5 A basic mediation model.

and indirect effects of T on Y are not separately identifiable when only T but not M is randomized (Pearl, 2001; Robins & Greenland, 1992).

In the mediation context, the potential outcomes framework is more complex because the mediator is also posited to be an outcome of the exposure T . The potential value of the mediator for participant i under the intervention condition is denoted $M_i(1)$, and the potential value of the mediator under the control condition is denoted $M_i(0)$. The potential values for the outcome are then expanded to include potential values of the mediator. Incorporating potential values for the mediators, the nested expression $Y_i(1, M_i(1))$ is the potential outcome if individual i receives the exposure and has the level of the mediator that would be obtained under the exposure, and $Y_i(0, M_i(0))$ is the potential outcome if individual i is in the control condition and has the level of the mediator that would be obtained under the control condition. As before, only one of these two nested potential outcomes is observed for each individual. There also exist two nested potential outcomes that can never be observed for any individual. These are $Y_i(0, M_i(1))$, the potential outcome if the individual is in the control group but has the potential value of the mediator they would have had under the intervention condition, and $Y_i(1, M_i(0))$, the potential outcome if the individual receives the intervention but has the potential value of the mediator that they would have had under the control condition.

The potential outcomes framework permits a variety of causal effects to be clearly defined; given identifying assumptions, they can then be estimated. Table 16.1 summarizes some of these

Table 16.1 Definitions of causal effects

<i>Effect</i>	<i>Definition</i>	<i>Comments</i>
Individual causal effect	$Y_i(1) - Y_i(0)$	In general, it is not possible to estimate this effect because only one of the potential outcomes is observed for each individual i .
Average causal effect (ACE)	$E[Y_i(1) - Y_i(0)]$	The expected difference between the potential outcomes in the population. This is the total effect of the treatment on the outcome and does not provide information about mechanisms (i.e., mediators) of the effect.
Natural direct effect	$E[Y_i(1, M_i(0)) - Y_i(0, M_i(0))]$	The effect of the intervention on the outcome, setting the mediator to the value that would have been obtained under the control condition.
Controlled direct effect	$E[Y_i(1, m) - Y_i(0, m)]$	The value of the mediator is set to a specific value m for individuals in both the treatment and control conditions.
Principal strata effect	$E[Y_i(1) - Y_i(0) \mid M_i(1) = 1, M_i(0) = 0]$	The effect of the treatment in an unobserved subgroup of individuals for whom the intervention has an effect on the mediator.
Natural indirect effect	$E[Y_i(1, M_i(1)) - Y_i(1, M_i(0))]$	The effect of the intervention on the outcome that arises from an intervention-induced change in the mediator.

Note: E is the expectancy operator.

effects, beginning with the individual causal effect and the ACE. Each of the effects may be of interest for a variety of questions about mechanisms by which T induces an effect on Y .

Specifying the causal effect of interest is important, particularly in the mediation context. By using the potential outcomes framework, it becomes obvious that the term *direct effect* (i.e., the effect of the exposure on the outcome that is not due to the mediator) is not specific enough. For example, we can define a causal effect, $E[Y_i(1, M_i(0)) - Y_i(0, M_i(0))]$: the causal effect of the intervention on the outcome, setting the mediator to the value that would have been obtained for the individual under the control condition. This causal effect has been referred to as the *pure direct effect* by Robins and Greenland (1992) and as the *natural direct effect* by Pearl (2001; see also Pearl, 2009). The rationale for calling this the natural direct effect is that this is the effect of the exposure on the outcome when setting the mediator to the value that it naturally would have been in the absence of the exposure. The rationale for calling this the pure direct effect is that if there is an interaction between the exposure and the mediator, then this effect does not include the interaction term; that is, it is the “pure” effect of the exposure on the outcome that is not due to the mediator.

Rather than setting the value of the mediator to one of its potential values, the researcher could choose to define a direct effect in which the value of the mediator is set to a particular value, m , of the researcher’s choosing for all individuals. This effect has been referred to as the controlled direct effect, $E[Y_i(1, m) - Y_i(0, m)]$. For example, if the mediator can take on values of 0 or 1 (i.e., it is a binary variable) then two controlled direct effects could be defined; one by setting $m = 0$ and the other by setting $m = 1$ for all individuals. We could also define an effect, $E[Y_i(1) - Y_i(0) \mid M_i(1) = 1, M_i(0) = 0]$, that has been referred to as the principal strata direct effect. This is the effect of the intervention on the outcome among the subpopulation for which the intervention had an effect on the mediator. Finally, we can define the natural indirect effect as $E[Y_i(1, M_i(1)) - Y_i(1, M_i(0))]$, which is the effect of the intervention on the outcome that is due to an intervention-induced change in the mediator. These effects are summarized in Table 16.1.

In summary, a distinction can be made among (a) the causal effect of an exposure on an outcome, setting the mediator to one of its potential values (e.g., $M_i(0)$), which differs across individuals; (b) a causal effect of an exposure on an outcome, setting the mediator to a specific value m for all individuals; (c) the causal effect of an exposure on an outcome among a latent (i.e., unobserved) subgroup of people for whom the exposure has an effect on mediator if they are exposed and it does not if they are not exposed; and (d) the causal effect of an exposure on an outcome that is due to an exposure induced change in the mediator. These various effects are further discussed in Albert (2008); Gallop et al. (2009); Lynch, Kerry, Gallop, and Ten Have (2008); Ten Have et al. (2007); Jo (2008); Sobel (2008); VanderWeele (2008); VanderWeele (2009a); VanderWeele and Vansteelandt (2009); Dunn and Bentall (2007); Emsley, Dunn, and White (2010); Holland (1988); and Imai, Keele, and Tingley (2010).

Post-treatment confounders as mediators

The situation becomes even more complex when a time-varying post-treatment confounder of the mediator and outcome is present and has itself been influenced by the exposure T . Use X_1 to denote this type of confounder.

As depicted in Figure 16.6, the total direct effect, partial direct effect, total indirect effect, or partial indirect effect may all be of interest depending on the scientific question under investigation. The total direct effect consists of all the paths that do not go through M , the mediator of scientific interest. In contrast, the partial direct effect consists only of the path between the exposure and the outcome that does not involve either the mediator or the post-treatment confounder. The total indirect effect consists of all pathways that go through M , including one

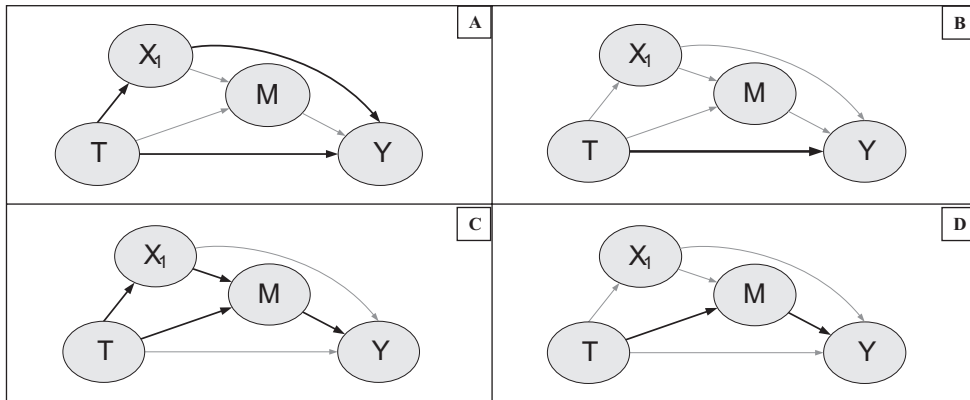


Figure 16.6 Mediation models incorporating a time-varying post-treatment covariate.

that goes through the post-treatment confounder. The partial indirect effect consists only of the pathway that goes through M but not through the post-treatment confounder.

The total direct effect and the total indirect effect (Figures 16.6, Panel A and 16.6, Panel C) are not identified and cannot be estimated without invoking additional assumptions (Avin, Shipster, & Pearl, 2005; Imai, Keele, & Tingley, 2010; Pearl, 2012; Robins, 2003). For example, they may be estimated using parametric assumptions such as no interactions between the exposure and the mediator (Robins, 2003).

The partial direct effect and partial indirect effect illustrated in Figures 16.6, Panel B and 16.6, Panel D, respectively, may be estimated using inverse probability weighting and are often of interest to substantive researchers (Coffman & Zhong, 2012), as illustrated by an application focused on mediation of a human immunodeficiency virus preventive intervention (Coffman & Kugler, 2012).

I will demonstrate their importance with an example. Suppose that individuals are randomly assigned to a substance-use treatment program (T) and that a component of the program encourages participants to change their social contexts (M ; e.g., who they spend time with and where they spend time); this, in turn, is hypothesized to affect the likelihood of relapse (Y). Suppose also that after assignment to the treatment, some of the participants changed employment (X_1). The change in employment may influence the individual's social context and likelihood of relapse. In addition, employment change may have been influenced by the treatment program even though this was not a goal of the treatment program. Thus, the researcher is interested in the effect $T \rightarrow M \rightarrow Y$ (Figure 16.6, Panel D) but must take into account the employment change. That is, the researcher is not interested in the effect, $T \rightarrow X_1 \rightarrow M \rightarrow Y$ (Figure 16.6, Panel B). Thus, it is important for researchers to clearly define the effect that they want to estimate.

For the reasons articulated above, it is important for researchers to first clearly articulate the scientific question(s) of interest, including the causal effect they are most interested in. Next, researchers should consider the plausibility of the assumptions needed to identify this causal effect for their specific data and research area. Sensitivity analyses for assessing the plausibility of the various assumptions have been proposed (e.g., Imai, Keele, & Yamamoto, 2010; Liu, Kuramoto, & Stuart, 2013; VanderWeele & Arah, 2011). Finally, researchers should choose a method that estimates the causal effect of interest. Sometimes it is not possible to identify the causal effect of interest under plausible assumptions. For example, if a researcher is interested in the indirect causal effect illustrated in Figure 16.6, Panel C and does not think it is plausible to assume, for example, no interactions between the exposure and mediator, it will be impossible to estimate the effect using currently available methods.

Treatment adherence and program fidelity

One specific type of mediator is adherence to or compliance with treatment assignment. Individuals who are randomly assigned to an intervention may not fully comply with the treatment protocol. Likewise, those assigned to the comparison condition may carry out aspects of the intervention, especially when contamination is possible or participants prefer to receive the intervention. In such situations, assignment to treatment or control conditions and exposure to treatment or control conditions are not the same. However, we would expect that individuals who are assigned to the treatment are more likely to comply with the treatment protocol than individuals assigned to the control condition. In other words, assigned treatment influences compliance, which in turn influences the outcome. However, researchers are generally not interested in the indirect effect in this context. Rather, they are generally interested in the causal effect of the treatment on the outcome among those who comply with treatment assignment. This effect is commonly referred to as the complier average causal effect (CACE; e.g., Little & Yau, 1998) and is analogous to the principal strata direct effect defined earlier in that it is the treatment effect among a latent subgroup of individuals who complied with treatment assignment (i.e., those who take the treatment if assigned to it and who do not take the treatment if not assigned to it). The subgroup is latent because principal strata are cross-classifications based on potential outcomes, some of which are not observed. The CACE is often estimated using *IV*s, which are further discussed in the next section.

Compliance with treatment can also be a dosage variable. For example, Crowley, Coffman, Feinberg, Greenberg, and Spoth (2013) examine the effect of high (more than four sessions) versus low (four or fewer sessions) dosage of a family program designed to prevent substance use among adolescents. In this case, some families assigned to the intervention received the treatment as designed, some self-selected not to fully complete the program, and others did not complete any of the program.

Program and/or implementation fidelity are similar issues to compliance. Complex or multicomponent interventions may not be carried out by the teachers or staff nurses as planned by the program developers, and, therefore, the participants are not exposed to the intervention to which they were randomly assigned. Program and/or implementation fidelity are then compromised and can be considered in the same manner as a compliance or adherence variable. In this situation, participants do not self-select into program fidelity, but nonetheless, there are likely to be confounders of program fidelity and subsequent outcomes. For example, consider a school-based prevention program designed to reduce the likelihood of adolescents initiating substance use. Some schools may be under-resourced, which leads to the teachers not implementing the entire intervention program. Students in these under-resourced schools may also have higher rates of substance use initiation. Note that Judd and Kenny (1981) regarded adherence and fidelity as mediators: intervention assignment has an effect on adherence or fidelity, which in turn has an effect on the outcome. Again, in this context, researchers are usually most interested in the effect of the intervention on the outcome among those for whom the intervention was implemented with some degree of fidelity rather than in the indirect effect.

Comparison of different types of intermediate variables

Program mediators, treatment implementation mediators, and post-treatment confounders are all influenced by treatment and are intermediate between the treatment and the outcome. However, they may be distinguished by the scientific question of interest and their role may vary from one

study to another. Program mediators are intentional targets of the intervention. For example, a researcher may design a substance-use prevention program to target a student's perceived norms about substance use as a mechanism for reducing substance use. At the same time, the researcher hopes that all students adhere to and receive the intervention to which they are assigned. Of course, interventions rarely go strictly according to plan; this results in treatment implementation mediators. Post-treatment confounders, variables that should be taken into account despite being of limited scientific interest, are usually not desirable but may nevertheless occur despite the researcher's best intentions.

The indirect effect is most often of interest in the case of program mediators, whereas the direct effect is most often of interest in the case of post-treatment confounders. For example, in a substance-use treatment program designed to prevent relapse (McCaffrey, Morral, Ridgeway, & Griffin, 2007), some individuals were institutionalized following randomization. These individuals would not have been able to use substances, or at least would have been less likely to use substances. In addition, the treatment may have influenced which individuals were institutionalized. Therefore, institutionalization mediated the effect of the treatment on relapse. McCaffrey et al. (2007) were interested in the direct effect rather than the indirect effect. That is, they wanted to know the effect of the treatment on relapse, controlling for or holding constant institutionalization as opposed to the effect of the treatment program on relapse that was due to institutionalization. In this case, the substantive findings will be quite different for these two effects.

Ultimately, all post-treatment variables (those affected by the treatment that in turn affect the outcome) are not scientifically equivalent. Depending on the type of post-treatment variable, different causal effect definitions and estimation methods may be called for. In addition, the effect in which the researcher is most interested, whether it be the direct or indirect effect or the treatment effect among a subgroup, may differ depending on the context and the scientific question to be addressed. In two different analyses of the same study or in two different studies, the same variable may sometimes be a mediator and other times a post-treatment confounder.

Instrumental variables

Instrumental variables (*IVs*) have an effect on the exposure *T* but unlike confounders, they do not have an effect on the outcome *Y* (see Figure 16.7). Because they do not have an effect on both the exposure and the outcome, failing to account for *IVs* will not result in biased treatment effects. *IV* estimators are frequently used to obtain unbiased estimates of the causal effects of non-randomized treatments. They are also frequently used in the presence of non-compliance for estimating the CACE. However, identification of variables that meet the criteria and assumptions of an *IV* can be challenging, and incorporation of *IVs* that are only weakly associated with treatment exposure can worsen the quality of the estimate of treatment effect (Bound, Jaeger, & Baker, 1995).

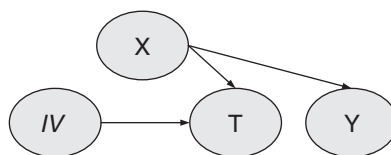


Figure 16.7 Instrumental variable effect, with confounder effect also shown.

The model given in Figure 16.7 can be written as a linear system of equations such that

$$Y = \beta_{Y0} + \beta_{Y1}X + \beta_{Y2}T + E_Y \quad (16.1a)$$

$$T = \beta_{T0} + \beta_{T1}X + \beta_{T2}IV + E_T \quad (16.1b)$$

Equation 16.1a shows the regression of the outcome, Y on the covariate, X and the treatment, T ; and Equation 16.1b shows the regression of the exposure T on the covariate X and the IV , including the error terms. To obtain the equation for prediction of the outcome from the covariate and the IV , the complete system can be solved by substituting Equation 16.1b into Equation 16.1a and rewriting:

$$Y = \beta_{Y0} + \beta_{Y1}X + \beta_{Y2}(\beta_{T0} + \beta_{T1}X + \beta_{T2}IV) + E_Y + E_T \quad (16.2)$$

$$= [\beta_{Y0} + \beta_{Y2}\beta_{T0}] + [\beta_{Y1}X + \beta_{Y2}\beta_{T1}X] + [\beta_{Y2}\beta_{T2}IV] + [E_Y + E_T]$$

$$= \beta_0 + \beta_1X + \beta_2IV + [E_Y + E_T],$$

where $\beta_0 = \beta_{Y0} + \beta_{Y2}\beta_{T0}$, $\beta_1 = \beta_{Y1} + \beta_{Y2}\beta_{T1}$, and $\beta_2 = \beta_{Y2}\beta_{T2}$.

The key coefficient of interest is β_{Y2} , which gives the effect of the treatment on the outcome. Using the IV approach, the estimate is obtained using β_2/β_{T2} from Equation 16.2. The IV estimate of β_{T2} is unbiased because, under the assumptions of the model (discussed next), there are no confounders of the IV and T . The estimate of β_2 is unbiased because, under the standard IV assumptions, there are no confounders of the IV and Y . Therefore, $\beta_{Y2} = \beta_2/\beta_{T2}$ is an unbiased estimate of the treatment effect.

IV estimators rely on strong assumptions. The first assumption is the exclusion restriction, which reflects the stipulation in the definition of an IV that the IV has an effect on the outcome only through its effect on the exposure. Another way to state this is that the IV should not be correlated with E_Y , the error in prediction of the outcome. The second assumption is that the relationship between the IV and T is not explained by any other covariates. That is, there are no confounders of the IV and T . The third assumption is that there are no confounders of IV and Y .

The problem with IV s is that they are rare and difficult to find (Hernan & Robins, 2006). Use of “weak” IV s that have a small correlation with the exposure can be more detrimental (i.e., result in estimates that are more biased) than an unadjusted estimate (Bound et al., 1995). In fact, Brookhart et al. (2006) showed that adjusting for a variable strongly related to the exposure and only weakly related to the outcome resulted in larger standard errors for the causal effect estimate. In addition, Pearl (2010) showed that in some cases, not only may this type of variable increase standard errors but it may also actually result in biased estimates of the causal effect. Thus, it is important that the IV be *strongly* related to the exposure and not related to the outcome. Good IV s are often created by policy changes, natural experiments (such as Mendelian randomization that occurs in genetics), some quasi-natural experiments, or regression discontinuity designs (Shadish et al., 2002). Even if a good IV is available to the researcher, IV estimation estimates the *local* ACE, which is the causal effect for a subgroup. For example, the CACE is a local ACE.

IV estimation for causal effects in the presence of non-compliance was described in Angrist, Imbens, and Rubin (1996). In the case of compliance, the IV estimator estimates the CACE. In compliance studies, the IV is randomization to treatment. Randomization to treatment

assignment is a plausible *IV* in compliance studies because the exclusion restriction is more plausible, and, due to randomization, it is plausible that there are no confounders of the *IV* and *T* and the *IV* and *Y*. For example, if the intervention is a blood pressure medication and the drug is not available to those who are assigned to the placebo condition, then the only way that the drug can have an effect on the outcome is if the participants take the drug. For many behavioral interventions, though, this assumption may not hold. For example, if the intervention is a weight-loss intervention that includes exercise, some individuals in the control condition may educate themselves and pursue strategies similar to those included in the intervention protocol. At the same time, not all in the intervention group will do the exercise. For good introductions to compliance issues, see Stuart, Perry, Le, and Ialongo (2008) and Little and Yau (1998).

Conclusions

In conclusion, third variables may play different roles depending on the scientific context. MacKinnon, Krull, and Lockwood (2000) discuss the equivalence of mediation and confounding, in that they both involve a difference in the relationship between the exposure and outcome after inclusion of a third variable. They further argue that mediation and confounding are distinguished conceptually rather than statistically. Momentarily disregarding the time ordering, if the arrow from the exposure to the mediator is reversed in Figure 16.5, the result is Figure 16.3, in which the mediator is now a confounder. Similarly, removing the arrow from the confounder to the outcome in Figure 16.3 results in Figure 16.7, in which the third variable is an *IV* rather than a confounder. The type of role that the third variable plays has implications for how it is modeled. For example, the researcher should adjust for common causes of the exposure and outcome but not for common effects of the exposure and outcome (Rosenbaum, 1984) or for *IV*s (Pearl, 2010).

Time-varying dynamics associated with exposures and moderators introduce even greater complexity. Time-varying moderators may be considered mediators that also moderate the effect of a later exposure on an even later outcome. A mediator, at a given time, can also moderate the effect of the exposure on the outcome at a later time or even the effect of an exposure on an outcome that occurs through a mediator at a later time.

To summarize, it is important for researchers to clearly state their research questions and to choose models that allow them to (a) estimate the exact causal effects that they are interested in and (b) take into account the effects they are not interested in. Judicious use of third variables can allow researchers to answer very specific research questions; however, failing to distinguish among the types of third variables can result in misleading and biased estimates of the effects of scientific interest.

Acknowledgments

I would like to thank the following individuals for valuable discussions that led to the ideas presented in this chapter: Daniel Almirall, Bethany Bray, Linda Collins, D. Max Crowley, John Graham, Mark Greenberg, Michael Hecht, Damon Jones, Stephanie Lanza, Julia Moore, Susan Murphy, Rob Nix, Wayne Osgood, C. J. Powers, Daniel Rivera, Joseph Schafer, Ed Smith, and Rob Turrisi. I would like to thank Amanda Applegate and Susan Henly for editorial assistance. This work was supported in part by a grant from the National Institute on Drug Abuse (NIDA) to Linda Collins: P50-DA010075-16. This work does not necessarily represent the official views of NIDA or the National Institutes of Health.

References

- Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in Medicine*, 27, 1282–1304.
- Almirall, D., McCaffrey, D. F., Ramchand, R., & Murphy, S. A. (2013). Subgroups analysis when treatment and moderators are time-varying. *Prevention Science*, 14, 169–178.
- Almirall, D., Ten Have, T., & Murphy, S. A. (2010). Structural nested mean models for assessing time-varying effect moderation. *Biometrics*, 66, 131–139.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 444–472.
- Avin, C., Shipster, I., & Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the International Joint Conferences on Artificial Intelligence* (pp. 357–363). Retrieved from www.ijcai.org/papers/0886.pdf
- Barber, J. S., Murphy, S. A., & Verbitsky, N. (2004). Adjusting for time-varying confounding in survival analysis. *Sociological Methodology*, 34, 163–192.
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90, 443–450.
- Bray, B. C., Almirall, D., Zimmerman, R. S., Lynam, D., & Murphy, S. A. (2006). Assessing the total effect of time-varying predictors in prevention research. *Prevention Science*, 7, 1–17.
- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163, 1149–1156.
- Coffman, D. L. (2011). Estimating causal effects in mediation analysis using propensity scores. *Structural Equation Modeling*, 18, 357–369.
- Coffman, D. L., Caldwell, L., & Smith, E. (2012). Introducing the at-risk average causal effect with application to HealthWise South Africa. *Prevention Science*, 13, 437–447.
- Coffman, D. L., & Kugler, K. C. (2012). Causal mediation of a human immunodeficiency virus preventive intervention. *Nursing Research*, 61, 224–230.
- Coffman, D. L., & Zhong, W. (2012). Assessing mediation using marginal structural models in the presence of confounding and moderation. *Psychological Methods*, 17, 642–664.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlational analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Crowley, D. M., Coffman, D. L., Feinberg, M., Greenberg, M. T., & Spoth, R. (2013). Evaluating the impact of implementation quality on evidence-based prevention programming: Methods for strengthening causal inference. *Prevention Science*. doi:10.1007/s11121-012-0352-8
- Dunn, G., & Bentall, R. (2007). Modelling treatment–effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). *Statistics in Medicine*, 26, 4719–4745.
- Emsley, R., Dunn, G., & White, I. R. (2010). Mediation and moderation of treatment effects in randomised controlled trials of complex treatments. *Statistical Methods in Medical Research*, 19, 237–270.
- Gallop, R., Small, D. S., Lin, J. Y., Elliott, M. R., Joffe, M. M., & Ten Have, T. R. (2009). Mediation analysis with principal stratification. *Statistics in Medicine*, 28, 1108–1130.
- Greenland, S., & Morgenstern, H. (2001). Confounding in health research. *Annual Review of Public Health*, 22, 189–212.
- Hernan, M. A., Hernandez-Diaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15, 615–625.
- Hernan, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist’s dream? *Epidemiology*, 17, 360–371.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–970.
- Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18, 449–484.
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15, 309–334.
- Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistics in Medicine*, 25, 51–71.
- Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods*, 13, 314–336.

- Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5, 602–619.
- Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, 27(2 Suppl.), S101–S108.
- Kraemer, H. C., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry*, 158, 848–856.
- Kraemer, H. C., Wilson, T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, 59, 877–883.
- Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health*, 21, 121–145.
- Little, R. J. A., & Yau, L. H. Y. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods*, 3, 147–159.
- Liu, W., Kuramoto, S. J., & Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science*, 14, 570–580.
- Lynch, K. G., Kerry, M., Gallop, R., & Ten Have, T. R. (2008). Causal mediation analyses for randomized trials. *Health Services Outcomes Research Methodology*, 8, 57–76.
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: LEA.
- MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding, and suppression effect. *Prevention Science*, 1, 173–181.
- McCaffrey, D. F., Morral, A. R., Ridgeway, G., & Griffin, B. A. (2007). Interpreting treatment effects when cases are institutionalized after treatment. *Drug and Alcohol Dependence*, 89, 126–138.
- Pearl, J. (2001). Direct and indirect effects. In P. Besnard & S. Hanks (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 411–420). San Francisco, CA: Morgan Kaufman.
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York, NY: Cambridge University Press.
- Pearl, J. (2010). On a class of bias-amplifying covariates that endanger effect estimates. In P. Grunwald & P. Spirtes (Eds.), *Proceedings of the Twenty-sixth Conference on Uncertainty in Artificial Intelligence* (pp. 417–424). Corvallis, OR: AUAI Press.
- Pearl, J. (2012). *Interpretable conditions for identifying direct and indirect effects* (Technical Report No. R-389). Los Angeles, CA: UCLA Cognitive Systems Laboratory.
- Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In P. Green, N. L. Hjort, & S. Richardson (Eds.), *Highly structured stochastic systems* (pp. 70–81). New York, NY: Oxford University Press.
- Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3, 143–155.
- Robins, J. M., Hernan, M. A., & Brumback, B. A. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550–560.
- Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A (General)*, 147, 656–666.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66, 688–701.
- Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31, 161–170.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100, 322–331.
- Schafer, J. L., & Kang, J. D. Y. (2008). Average causal effects from non-randomized studies: A practical guide and simulated example. *Psychological Methods*, 13, 279–313.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Sobel, M. E. (1995). Causal inference in the social and behavioral sciences. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 1–38). New York, NY: Plenum.
- Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, 33, 230–251.
- Stuart, E. A., Perry, D. F., Le, H.-N., & Ialongo, N. S. (2008). Estimating intervention effects of prevention programs: Accounting for noncompliance. *Prevention Science*, 9, 288–298.

- Ten Have, T. R., & Joffe, M. M. (2012). A review of causal estimation of effects in mediation analysis. *Statistical Methods in Medical Research*, 21, 77–107.
- Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., & Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics*, 36, 926–934.
- VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters*, 78, 2957–2962.
- VanderWeele, T. J. (2009a). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20, 18–26.
- VanderWeele, T. J. (2009b). On the distinction between interaction and effect modification. *Epidemiology*, 20, 863–871.
- VanderWeele, T. J., & Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments and confounders. *Epidemiology*, 22, 42–52.
- VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2, 457–468.
- Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, 25, 650–707.

JOINT MODELS FOR LONGITUDINAL DATA AND TIME-TO-EVENT OCCURRENCE

Yue Liu and Lei Liu

In longitudinal studies in the biomedical and health sciences, disease progress is often monitored by a marker over time, e.g., CD4 count in AIDS studies, hemoglobin level in end-stage renal disease (ESRD) patients. There could exist an event of interest, e.g., death or diagnosis of a specific disease. Patients may also drop out of the study due to side effects, ineffective treatment, or poor health status (too sick to continue the study). These events may be correlated with the longitudinal markers. The relation between the longitudinal data and time to event data is often of interest. For example, if the survival outcome takes a long time to occur, we might be interested in using a biomarker as a surrogate to the survival endpoint. On the other hand, if our primary interest is the longitudinal outcome, we would like to account for the possible informative drop-out to reduce potential bias.

A classical approach to evaluating the effect of longitudinal biomarkers on time to event is to use observed values as a time-varying covariate in a survival model for the event time (e.g., Cox proportional hazards model; Cox 1972). In the kidney disease setting, Robinson et al. (2005) and Bradbury et al. (2008) applied this approach to the effect of time-varying hemoglobin measures on survival. However, there are several drawbacks to the method. First, it requires a complete set of repeated measures in a time-continuous process. In reality, the biomarker (e.g., hemoglobin) is measured only at discrete time points, which may not be observable at the time of event occurrence. Although the value of the biomarker at event time can be obtained by some imputation techniques (e.g., last observation carried forward), this practice could be crude and lead to inappropriate inferences, especially when the time interval between adjacent observations is long. Second, patient survival to event occurrence might depend on the “underlying true” (or expected) values of biomarkers, rather than the observed values accompanying measurement errors. In this situation, the estimated parameters from such a model are biased toward the null (Prentice, 1982). Third, the hazard rate may depend largely on the change and variability in the repeated measures. As Regidor et al. (2006) found out, a decrease in hemoglobin over time was associated with a higher death risk, independent of the baseline hemoglobin value. However, fitting a Cox model with time-varying hemoglobin values cannot capture such features.

A solution is to use the two-stage model to first obtain an estimated trajectory of the biomarker for each subject during the follow-up period, and then incorporate such measures into the survival model. In the first stage, Dafni and Tsiatis (1998) used a mixed effects model to

describe the progression of the time-varying marker. The marker value at each event time can be calculated by the empirical Bayes estimate from the mixed model, based on the observed history of individuals who did not have an event up to that time. The empirical Bayes estimates, i.e., the posterior expected value of random effects conditional on the observed data, are then plugged in the survival model as predictors for time to event. Such a method addresses the afore-mentioned concerns. However, it may suffer from selection bias: patients with poorer health conditions have worse biomarker values (e.g., lower CD4), as well as a higher mortality rate or drop-out rate (thus shorter follow-up). Consequently, worse biomarker values would be observed less often; for example, the lower values of CD4 would be under-represented and the higher values over-represented, resulting in a selection bias.

In view of this, the disease progress model and the survival model should be developed jointly, since each process contains information on the other. Joint random effects models have been considered an appealing tool to describe longitudinal and time-to-event data simultaneously. Selection bias can be accounted for by the random effects shared between the mixed model of longitudinal markers and survival model for time to event. Estimation is carried out by maximizing the joint likelihood of the two correlated processes, rather than separately from each of the two stages in the two-stage method.

Earlier work in joint modeling includes Wu and Carroll (1988), Schluchter (1992), De Gruttola and Tu (1994), Faucett and Thomas (1996), and Wulfsohn and Tsiatis (1997). Among these papers, a linear mixed model was assumed for repeated measures. However, different assumptions were made for the survival outcome; e.g., lognormal (De Gruttola & Tu, 1994; Schluchter, 1992), probit for discrete time points (Wu & Carroll, 1988), piecewise constant baseline hazard (Faucett & Thomas, 1996), or a semiparametric Cox model with unspecified baseline hazard. Henderson et al. (2000) formulated a class of models for the joint analysis of longitudinal measures of psychiatric disorder and an associated event. More recent works include Tsiatis and Davidian (2001, 2004); Xu and Zeger (2001); Lin et al. (2002); Ratcliffe, Guo, and Ten Have (2004); Yu et al. (2004); Hsieh, Tseng and Wang (2006); Vonesh, Greene and Schluchter (2006), M. Liu and Ying (2007), Ding and Wang (2008); Ye, Lin and Taylor (2008), L. Liu, Wolfe, & Kalbfleisch, (2007); L. Liu, Huang, & O'Quigley (2008); L. Liu Ma, & O'Quigley (2008); L. Liu & Huang (2009); and Ghosh and Tu (2008), among many others.

Purpose

Although shared random effects models have been well discussed in the statistical literature, they have not been widely adopted in practical analysis in the health sciences. In this chapter, we will introduce this type of models to investigators in health research, and show how to implement these methods in the real world. The rest of the paper is organized as follows. First, we define the basic notation and likelihood for a simple joint model. Then, we describe the estimation procedure. We present the implementation method in SAS and interpret the results with the application to a dataset from an AIDS study. We close with some concluding remarks and possible future research interests.

Joint model

Suppose we record a repeated measure Y_{ij} at the j th observation time t_{ij} for subject i . There is a parallel correlated event D_i , e.g., death or informative drop-out. There also exists an independent censoring event C_i , which is not related to the repeated measures (e.g., moving out of the region). Define by $X_i = \min(C_i, D_i)$ the follow up time, and $\Delta_i = I(D_i \leq C_i)$ the event indicator. Denote

the hazard for the failure time D_i by $\lambda_i(t)$. Wulfsohn and Tsiatis (1997) proposed a shared random effects model of CD4 counts and time-to-death where:

$$y_{ij} = Z_{ij}^T \beta + a_i + b_i t_{ij} + e_{ij} \quad (17.1)$$

and

$$\lambda_i(t) = \lambda_0(t) \exp\left(W_i(t)^T \alpha + \gamma(a_i + b_i t)\right), \quad (17.2)$$

where a_i and b_i are random intercept and slope in the mixed model for repeated measures. Denote by Z_{ij} and $W_i(t)$ (possible time dependent) covariate vectors for repeated measures and failure time, respectively, with coefficients β and α . The baseline hazard for death is denoted by $\lambda_0(t)$. In Equation 17.2, a Cox proportional hazards model is used to describe death time, while the death hazard depends on the “underlying value” of repeated measures $a_i + b_i t$ through the random effects. The magnitude of the association between longitudinal and survival outcome is denoted by coefficient γ . If covariate vectors Z and W are disjoint, Equation 17.2 is sometimes written as

$$\lambda_i(t) = \lambda_0(t) \exp\left(W_i(t)^T \alpha + \gamma\left(Z_i(t)^T \beta + a_i + b_i t\right)\right). \quad (17.3)$$

In Equation 17.3, we can describe the direct effect of the “expected value” of repeated measures ($Z_i(t)^T \beta + a_i + b_i t$) on the time-to-event. Note here we use $Z_i(t)$ to denote the covariate vector for repeated measures at time t .

Assume a normal random error $e_{ij} \sim^{iid} N(0, \sigma_e^2)$. Denote by $\pi_\theta(a_i, b_i)$ the random effects density, where θ is the parameter, e.g., covariance matrix for correlated random effects (a_i, b_i) . Normal random effects are most often assumed due to the ease of implementation, although other distributions can be taken as well (L. Liu & Yu, 2008; Rizopoulos et al., 2008; Tsonaka et al., 2009). The likelihood for joint model (1) and (2) can be written as:

$$L_i = \int \prod_{j=1}^{n_i} \frac{1}{(\sqrt{2\pi}\sigma_e)^{n_i}} \exp\left[-\frac{1}{2\sigma_e^2} \sum_{j=1}^{n_i} e_{ij}^2\right] [\lambda_0(x_i) \exp\left(W_i(t)^T \alpha + \gamma(a_i + b_i t)\right)]^{\Delta_i} \exp\left[-\int_0^{x_i} \lambda_0(t) \exp\left(W_i(t)^T \alpha + \gamma(a_i + b_i t)\right) dt\right] \pi_\theta(a_i, b_i) da_i db_i, \quad (17.4)$$

where x_i is the realization of the follow-up time X_i , n_i is number of observed repeated measures, and $e_{ij} = y_{ij} - Z_{ij}^T \beta - a_i - b_i t_{ij}$.

An alternative formulation of joint models assumes that time-to-event depends on the random parameters of a subject’s underlying response profile (e.g., random intercept and slope) separately (Guo & Carlin, 2004; Ribaudo et al., 2000; Vonesh et al., 2006). The model is

$$\lambda_i(t) = \lambda_0(t) \exp\left(W_i(t)^T \alpha + \gamma_1 a_i + \gamma_2 b_i\right). \quad (17.5)$$

By this formulation we can also examine the effect of covariates (e.g., treatment) on both repeated measures and time-to-event in Equations 17.1 and 17.4, respectively, after adjusting for other risk factors and random effects.

For better illustration, hereafter we call the joint model with Equations 17.1 and 17.2 “model I,” while that with Equations 17.1 and 17.4 are referred to as “model II.” Model I is more appealing

when we are interested in the prediction of survival outcome by the longitudinal biomarker, while model II is used more often in sensitivity analysis of the impact of informative drop-out on the longitudinal outcome.

When only a random intercept is present, model I and model II are equivalent. However, when both random intercept and slope are present, model I is not nested within model II, and vice versa. As a result, the likelihood ratio test cannot be used to compare the models. Instead, in this paper we will use Akaike Information Criteria (AIC) to compare the non-nested model I and model II, using an example with a real dataset (below).

Note. We would like to make a clear distinction between different event processes, which could affect the repeated measures differently. For example, an informative drop-out event stops the follow-up, so that the subsequent measurements are not observable. However, such a drop-out event does not change the underlying course of the longitudinal process. This is different from a terminal event such as death, which changes the repeated measures process afterwards. For example, death precludes further accumulation of medical cost, so the repeated measures of medical cost (e.g., monthly medical cost) are zero after death. For such an outcome, we observe the whole repeated measure process, even after the terminal event. In this situation, repeated measures and survival are not independent, even after conditioning on random effects. On the other hand, for outcomes such as CD4 count or blood pressure, there is no real clinical interest in any definition of such biomarkers after death. Ribaldo et al. (2000), Pauler et al. (2003), and Kurland and Heagerty (2005) also distinguished drop-out and death explicitly.

Estimation

A notable hurdle in the application of joint random effects models lies in the difficulty in the implementation of the estimation process. The estimation often involves the EM algorithm (Dempster et al. 1977) with a Monte Carlo (or quadrature) procedure in the E-step to approximate the conditional expectation terms of random effects given observed data, which do not have a closed form. Such programs are challenging in implementation and slow in convergence (L. Liu, Wolfe, & Kalbfleisch, 2007). However, substantial progress has been made recently to overcome this hurdle. Available software has been developed in R function `jointModel` of the `JM` package (Rizopoulos, 2008). On the other hand, Guo and Carlin (2004) and Vonesh et al. (2006) proposed an adaptive Gaussian quadrature estimation method, which can be conveniently implemented in SAS Proc NLMIXED (Littell et al., 2006). An advantage of such method is that it can be easily adapted to more complicated settings, see L. Liu, Huang, & O'Quigley (2008) and L. Liu, Ma, & O'Quigley (2008) and L. Liu (2009). An alternative estimation method is implemented in a freely available software `aML` (Lillard & Panis, 2003), using non-adaptive Gaussian quadrature. The strength of `aML` is that it can handle multi-level repeated measures and survival data (L. Liu et al., 2008a). However, caution must be exercised in that a higher number of quadrature points are needed to obtain reliable estimates for non-adaptive Gaussian quadrature (Lesaffre & Spiessens, 2001).

In this chapter, we will use SAS Proc NLMIXED to implement our method because it is convenient to implement, estimation is accurate, and easy to extend to more complicated models. SAS Proc NLMIXED allows users to construct the log likelihood function (in the integrand of Equation 17.4 with respect to random effects) by regular SAS statements, similar to those in the DATA step. Parameters are estimated through maximizing the likelihood function (Equation 17.4) by using the option “general” in the “model” statement in Proc NLMIXED. Specifically, we assume the baseline hazard in the Cox model is piecewise constant. L. Liu and Huang (2008) introduced the implementation of the piecewise constant baseline hazard in SAS Proc NLMIXED. The sample SAS code is given in the Appendix.

Data analysis

In this section we apply our method to an AIDS dataset. L. Liu and Huang (2009) analyzed the repeated measures of CD4 count and survival jointly in the Terry Bein Community Programs for Clinical Research on AIDS (CPCRA) data (Abrams et al., 1994; Neaton et al., 1994). The study number is CPCRA 002 and clinical trial ID is NCT00000969 on clinicaltrials.gov website. This study enrolled 467 patients: 230 were randomized to receive didanosine (ddI) and the others received zalcitabine (ddC). During the follow-up (median: 13 months, range: 1–21 months), there were 100 and 88 deaths in the didanosine and zalcitabine groups, respectively. Repeated measures CD4 values for each patient were taken at baseline and every two months up to 20 months. There are a large number of intermittent missing observations of CD4, which are assumed to be missing at random (Little & Rubin, 2002). We are interested in the relation between CD4 repeated measures and the time to death. To address the right skewness and heteroscedasticity of CD4 measures (mean: 74; median: 30), and a small portion (2.3%) of zero CD4 values, we take $\log(\text{CD4}+1)$ as the outcome. We plot the CD4 count over time in Figure 17.1. Panels A and B in Figure 17.1 show that patients with shorter follow-up were associated with a lower intercept of CD4, manifesting the association between CD4 values and time to death.

Five covariates are incorporated in both repeated measures and survival models: treatment (Trt; $\text{Trt1} = \text{ddC}$, $\text{Trt0} = \text{ddI}$), gender ($1 = \text{female}$, $0 = \text{male}$), previous opportunistic infection (PrevOI; ($1 = \text{AIDS diagnosis at baseline}$, $0 = \text{no AIDS diagnosis}$), stratum ($1 = \text{AZT intolerance}$, $0 = \text{AZT failure}$), and baseline hemoglobin (Base Hb; centered at mean = 12). Time since study entry (in years) is included as a covariate in the model for repeated CD4 measures. We fit the joint model by the adaptive Gaussian quadrature method, assuming a piecewise constant baseline hazard (with 10 nodes) for death. The implementation can be conveniently carried out in SAS Proc NLMIXED. Of note, in the SAS code for model I, we take $a_i + b_i t$ as constant (valued at the median time) within each quantile interval for ease of implementation. The results from model I and model II are given in Table 17.1.

Between these two models, most parameter estimates in the repeated measures part are similar. However, there is noticeable distinctness for parameters in the survival part and variance

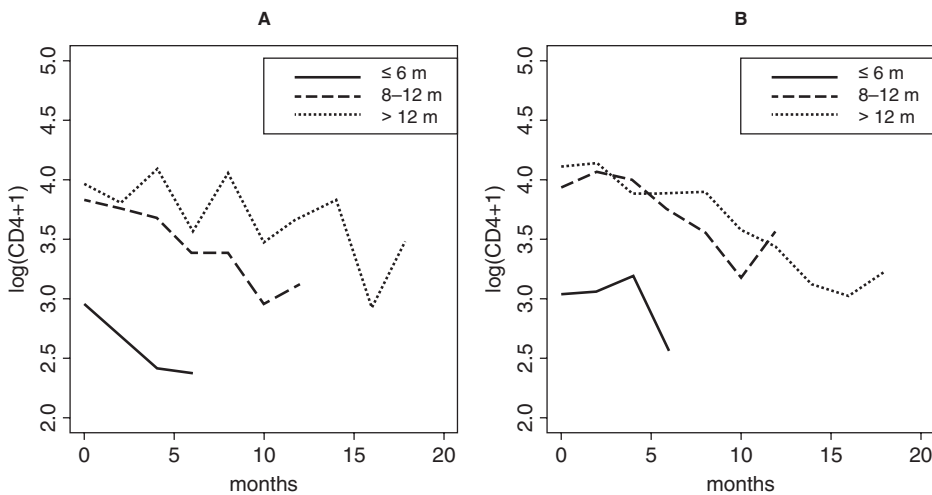


Figure 17.1 Exploratory plots of CD4 repeated measures. Panel A shows mean CD4 count over time in the ddC group; panel B shows mean CD4 count over time in the ddI group. The plots in panels A and B have different lengths of follow up: ≤ 6 months, and > 12 months. From L. Liu and Huang (2009). Reprinted with permission.

Table 17.1 Joint analysis of CD4 count and survival

	<i>Model I</i>			<i>Model II</i>		
	<i>Estimate</i>	<i>SE</i>	<i>p-value</i>	<i>Estimate</i>	<i>SE</i>	<i>p-value</i>
CD4 Value						
Intercept	4.082	0.173	< .0001	4.077	0.175	< .0001
Treatment	−0.099	0.107	.36	−0.101	0.107	.35
Time (years)	−0.773	0.064	< .0001	−0.758	0.068	< .0001
Female	0.256	0.187	.17	0.249	0.188	.19
Previous OI	−0.938	0.142	< .0001	−0.945	0.143	< .0001
Stratum	0.111	0.132	.40	0.117	0.133	.38
Base Hb	0.230	0.032	< .0001	0.232	0.032	< .0001
σ_c^2	0.305	0.016		0.299	0.170	
Survival						
Treatment	−0.294	0.153	.05	−0.307	0.158	.05
Female	0.032	0.258	.90	−0.111	0.268	.68
Previous OI	0.985	0.239	< .0001	1.087	0.244	< .0001
Stratum	−0.137	0.169	.42	−0.122	0.176	.49
Base Hb	−0.354	0.052	< .0001	−0.399	0.054	< .0001
γ	−0.058	0.021	.007			
γ_1				−0.504	0.077	< .0001
γ_2				−0.320	0.187	.09
Variance Components						
σ_a^2	1.167	0.090	< .0001	1.214	0.092	< .0001
σ_b^2	0.577	0.102	< .0001	0.630	0.101	< .0001
σ_{ab}	0.030	0.077	.70	−0.054	0.073	.46

Note: OI = opportunistic infection. Hb = hemoglobin. From L. Liu and Huang (2009). Reprinted with permission.

components. In model I, the estimate for γ is -0.058 ($p = .007$), suggesting the negative association between the hypothetical “underlying value” of CD4 and death hazard, i.e., a higher underlying value of CD4 is associated with a lower mortality. In model II, both $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are significantly (or marginally significantly) less than 0, implying higher initial values and a slower drop in CD4 counts is associated with a better survival. The AIC for model I vs. model II is 2,725.3 vs. 2,688.9 (smaller is better). Therefore, the hazard of death is more likely dependent on the initial level and slope of CD4 count in different magnitudes, rather than dependent on the hypothetical “underlying value” of CD4. Specifically, after adjusting for other risk factors, a 10% decrease in CD4 value (or 0.1 unit decrease in log CD4 value) at baseline is associated with 5% ($= e^{0.050} - 1$) increase in the death hazard, while a 10% drop of CD4 value per year is associated with 3% increase in the death hazard.

In model II, we can see no significant difference between ddC and ddI in CD4 values. However, the treatment effect is significant for death hazard, as ddC reduces the mortality rate by 26% ($p = .05$). There is a highly significant linearly decreasing trend in log CD4 values over time,

while the CD4 in the original scale decreases at a rate of 53% each year. PrevOI and baseline hemoglobin level are highly significant for both longitudinal and survival outcomes (all $p < .0001$). Patients with previous opportunistic infection (AIDS diagnosis) at baseline had 61% lower CD4 values and 3.0 times higher death hazard. One unit lower hemoglobin level at baseline is associated with 21% lower CD4 values over time and 49% higher death rate.

From the variance components, we note that the estimates of σ_a^2 and σ_b^2 are highly significant, indicating the presence of heterogeneity in terms of both the random intercept and the random slope for repeated measures. However, the covariance structure between the random effects a_i and b_i shows no significance.

As an example to assess the fit of the preferred model II, we show the residual plots in Figure 17.2. The residuals can be obtained directly by “predict” statement in SAS Proc NLMIXED, with empirical Bayes (EB) estimates for the random effects. The parameter estimates from Equation 17.5 can be used to calculate the Martingale residuals (Therneau, Grambsch, & Fleming, 1990), in order to evaluate the model adequacy for the survival outcome. Martingale residuals in the Cox proportional hazards model are defined as

$$\Delta_i - \int_0^t Y_i(s) \exp\left(W_i(s)^T \hat{\alpha} + \gamma_1 \hat{a}_i + \gamma_2 \hat{b}_i\right) d\hat{\Lambda}_0(s), \quad (17.6)$$

where $Y_i(s)$ indicates that if subject i is at risk at time s , and $\Lambda_0(s)$ is the estimated cumulative baseline hazard function. More model diagnostic methods for joint longitudinal and time-to-event data can be found in Dobson and Henderson (2003)

In Figure 17.2, we show in Panels A and B the residuals from Equation 17.1 and martingale residuals from Equation 17.5 vs. treatment, respectively. In Panel C, we show the residual plot for CD4 counts vs. time (in years). In Panel D, we show the martingale residual plot for baseline

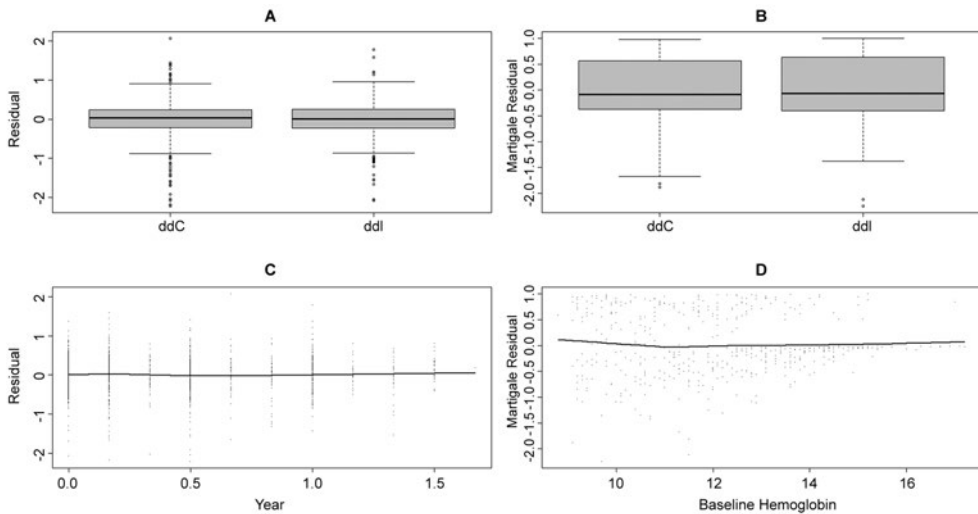


Figure 17.2 Residual plots. (A) Box plot of residuals for log(CD4+1), ddl vs. ddC; (B) Box plot of martingale residuals for death, ddl vs. ddC; (C) Residuals for log(CD4+1) over time (in years); (D) Martingale residuals for death over baseline hemoglobin. Lowest estimate is given in panels (C) and (D) to show the pattern.

hemoglobin. In both Panels C and D, Lowess fit is used to describe the possible non-linear trend. No clear pattern can be found in any of these residual plots, suggesting a reasonably good fit of our model.

Discussion

Although the study of joint random effects models has gained increasing interest among statisticians, the application of such methods in real data analysis in health research is still very limited (Do, 2002). However, the validity and robustness of the clinical investigations often entail appropriate sensitivity analyses for missing data, making such joint model framework very appealing in medical research. For example, we were among the first to apply a joint random effects model (model II) to the alcohol research field, investigating the relationship between drop-out time and longitudinal drinking outcomes (Johnson et al., 2011). We found that after adjusting for risk factors, the time to drop-out did not depend on the random effects from the longitudinal model of drinking outcomes, suggesting that drop-out was not informative. Furthermore, the parameter estimates for the longitudinal outcome from this joint model were similar to those in the mixed model for longitudinal data only, which assumed that data were missing at random. Our analysis provided evidence of the validity of the missing-at-random assumption. It was well received, commented by a reviewer as “very instructive, and convincing.”

Various generalizations of joint models are available. For example, Brown, Ibrahim, and De Gruttola (2005) proposed a joint model with a non-parametric covariate effect (specified by a cubic B-spline) for the longitudinal markers. On the other hand, the longitudinal markers might have a non-linear effect on time-to-event. For example, Ma, Ebben, Xia, and Collins (1999) found that hematocrit had a non-linear effect on patient’s survival in ESRD patients. A model can be thus proposed as

$$\lambda_i(t) = \lambda_0(t) \exp(W_i(t)^T \alpha + f(E(y_{it}))), \quad (17.7)$$

where $f(\bullet)$ is an unknown function whose functional form is of interest. We can thus describe the accurate relation between expected value of longitudinal markers and survival.

In addition to the dependence on the underlying (expected) value of the biomarker, death hazard could also depend on the variability across time in repeated measures. For example, Yang et al. (2007) studied the impact of the patient-level hemoglobin trajectory in a base period on death hazard using a two-stage model. They showed that death is associated with the absolute level (intercept), temporal trend (slope), and deviation from linear trend denoted by the *residual standard deviation* in the linear model for each individual’s trajectory. It is of interest to incorporate such variability in a joint model of repeated measures and survival. In another situation, the National Kidney Foundation (NKF) recommended a target range for hemoglobin of 11 to 12 g/dL (Hct 33–36%) for ESRD patients. There may be adverse effect if either below or above this level. Thus, a joint model of interest would involve the average of (absolute) differences between the observed and the target values. Denote by D_u and D_l the average absolute deviation from the upper and lower limits, respectively. A new model could be proposed as:

$$\lambda_i(t) = \lambda_0(t) \exp\left(W_i(t)^T \alpha + \gamma_1 a_i + \gamma_2 b_i + \gamma_3 D_u + \gamma_4 D_l\right). \quad (17.8)$$

Coefficients γ_3 and γ_4 denote how a one-unit change above or below the target range can affect patient survival. In this case, death hazard depends on observed values (represented by D_u

and D_j) in addition to random effects. However, random effects a_i and b_i are correlated with the observed values. Caution should be exercised for the interpretation of covariate coefficients in this model.

We have worked on several new methods in the joint modeling framework. For example, a common problem with longitudinal data is that repeated measures are often taken at informative observation times, i.e., repeated measures and observational times are correlated. For example, patients in poorer health are more likely to seek medical treatment and their medical cost for each visit tends to be higher. Such data are sometimes called “recurrent marker data” in that the repeated measures (e.g., medical costs) are marker of the recurrent events (e.g., hospital visits) process. L. Liu, Huang, and O’Quigley (2008) proposed a joint random effects model of repeated measures in the presence of both informative observation times and a dependent terminal event. They added an additional component for recurrent observational process in the joint model of repeated measures and survival. Other work on this topic includes Sun, Sun, and D. Liu (2007).

Further, we can apply the latent class framework into joint modeling to identify underlying subgroups of patients with similar behaviors from longitudinal and survival processes (Y. Liu, 2013). Assuming there exist K latent classes in the population, we first model the chance of belonging to different latent classes for each subject in a multinomial logit regression. Then under each latent class, we have a joint random effects model of longitudinal and survival outcomes. Class specific parameters for regression coefficients, baseline hazards, random effects, and error terms can be used in different classes to characterize the heterogeneity across latent classes. Estimation is obtained by maximizing the log full likelihood function. The optimal number of latent groups K can be determined by comparing BIC values for a series of models with varying specifications of the number of groups.

Extension to multivariate processes is currently an active research area. For example, L. Liu and Huang (2009) used a joint model to study the repeated measures CD4 and recurrent opportunistic diseases, with a terminal event death in the CPCRA data set. Some of the most up-to-date research was summarized in a workshop on the statistical analysis of multi-outcome data (available at www.lsta.upmc.fr/SAM2012/Program.html).

Acknowledgements

This research was partly supported by AHRQ grant R01 HS020263.

References

- Abrams, D. I., Goldman, A. I., Launer, C., Korvick, J. A., Neaton, J. D., Crane, L. R., . . . Terry Beirn Community Programs for Clinical Research on AIDS. (1994). A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. *New England Journal of Medicine*, 330, 657–662.
- Bradbury, B. D., Wang, O., Critchlow, C. W., Rothman, K. J., Heagerty, P., Keen, M., & Acquavella, J. F. (2008). Exploring relative mortality and epoetin alfa dose among hemodialysis patients. *American Journal of Kidney Disease*, 51, 62–70.
- Brown, E. R., Ibrahim, J. G., & De Gruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61, 64–73.
- Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187–220.
- Dafni, U. G., & Tsiatis, A. A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 54, 1445–1462.
- De Gruttola, V., & Tu, X. M. (1994). Modeling progression of CD4-Lymphocyte count and its relationship to survival time. *Biometrics*, 50, 1003–1014.

- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Ding, J., & Wang, J.-L. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*, 64, 546–556.
- Do, K.-A. (2002). Biostatistical approaches for modeling longitudinal and event time data. *Clinical Cancer Research*, 8, 2473–2474.
- Dobson, A., & Henderson, R. (2003). Diagnostics for joint longitudinal and dropout time modeling. *Biometrics*, 59, 741–751.
- Faucett, C. L., & Thomas, D. C. (1996). Simultaneously modeling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine*, 15, 1663–1685.
- Ghosh, P., & Tu, W. (2008). Assessing sexual attitudes and behaviors of young women: A joint model with nonlinear time effects, time varying covariates, and dropouts. *Journal of the American Statistical Association*, 103, 1496–1507.
- Guo, X., & Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *American Statistical Association*, 58, 16–24.
- Henderson, R., Diggle, P., & Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics*, 1, 465–480.
- Hsieh, F., Tseng, Y. K., & Wang, J. L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, 62, 1037–1043.
- Johnson, B. A., Ait-Daoud, N., Seneviratne, C., Roache, J. D., Javors, M. A., Wang, X.-Q., . . . Li, M. D. (2011). Pharmacogenetic approach at the serotonin transporter gene as a method of reducing the severity of alcohol drinking. *American Journal of Psychiatry*, 168, 265–275.
- Kurland, B. F., & Heagerty, P. J. (2005). Directly parameterized regression conditioning on being alive: Analysis of longitudinal data truncated by deaths. *Biostatistics*, 6, 241–258.
- Lesaffre, E., & Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Applied Statistics*, 50, 325–335.
- Lillard, L. A., & Panis, C. W. A. (2003). aML multilevel multiprocess statistical software, Version 2.0. Econ Ware, Los Angeles, CA. Freely available at www.applied-ml.com
- Lin, H., Turnbull, B. W., McCulloch, C. E., & Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97, 53–65.
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed model* (2nd ed.). Cary, NC: SAS Institute.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.
- Liu, L. (2009). Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Statistics in Medicine*, 28, 972–986.
- Liu, L., & Huang, X. (2008). The use of Gaussian quadrature for estimation in frailty proportional hazards models. *Statistics in Medicine*, 27, 2665–2683.
- Liu, L., & Huang, X. (2009). Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Applied Statistics*, 58, 65–81.
- Liu, L., Huang, X., & O'Quigley, J. (2008). Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*, 64, 950–958.
- Liu, L., Ma, J. Z., & O'Quigley, J. (2008). Joint analysis of multi-level repeated measures data and survival: An application to the end stage renal disease (ESRD) data. *Statistics in Medicine*, 27, 5679–5691.
- Liu, L., Wolfe, R. A., & Kalbfleisch, J. D. (2007). A shared random effects model for censored medical costs and mortality. *Statistics in Medicine*, 26, 139–155.
- Liu, L., & Yu, Z. (2008). A likelihood reformulation method in non-normal random effects models. *Statistics in Medicine*, 27, 3105–3124.
- Liu, M., & Ying, Z. (2007). Joint analysis of longitudinal data with informative right censoring. *Biometrics*, 63, 363–371.
- Liu, Y. (2013) *Joint latent class modeling in longitudinal and survival processes*. (Doctoral thesis). University of Virginia, Charlottesville. Manuscript in preparation.
- Ma, J. Z., Ebben, J., Xia, H., & Collins, A. J. (1999). Hematocrit level and associated mortality in hemodialysis patients. *Journal of the American Society of Nephrology*, 10, 610–619.
- Neaton, J. D., Wentworth, D. N., Rame, F., Hogan, C., Abrams, D. I., & Deyton, L. (1994). Considerations in choice of a clinical endpoint for AIDS clinical trials. *Statistics in Medicine*, 13, 2107–2125.

- Pauler, D.K., McCoy, S., & Moinpour, C. (2003). Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Statistics in Medicine*, 22, 795–809.
- Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69, 331–342.
- Ratcliffe, S. J., Guo, W., & Ten Have, T. R. (2004). Joint modeling of longitudinal and survival data via a common frailty. *Biometrics*, 60, 892–899.
- Regidor, D.L., Kopple, J.D., Kovesdy, C.P., Kilpatrick, R.D., McAllister, C. J., Aronovitz, J., . . . Kalantar-Zadeh, K. (2006). Associations between changes in hemoglobin and administered erythropoiesis-stimulating agent and survival in hemodialysis patients. *Journal of the American Society of Nephrology*, 17, 1181–1191.
- Ribaudo, H. J., Thompson, S. G., & Allen-Mersh, T. G. (2000). A joint analysis of quality of life and survival using a random effect selection model. *Statistics in Medicine*, 19, 3237–3250.
- Rizopoulos, D. (2008). The JM package. Available at <http://cran.rproject.org/web/packages/JM/JM.pdf>
- Rizopoulos, D., Verbeke, G., Lesaffre, E., & Vanrenterghem, Y. (2008). A two-part joint model for the analysis of survival and longitudinal binary data with excess zeros. *Biometrics*, 64, 611–619.
- Robinson, B. M., Joffe, M. M., Berns, J. S., Pisoni, R. L., Port, F. K., & Feldman, H. I. (2005). Anemia and mortality in hemodialysis patients: Accounting for morbidity and treatment variables updated over time. *Kidney International*, 68, 2323–2330.
- Schluchter, M. D. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine*, 11, 1861–1870.
- Sun, J., Sun, L., & Liu, D. (2007). Regression analysis of longitudinal data in the presence of informative observation and censoring times. *Journal of the American Statistical Association*, 102, 1397–1406.
- Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77, 147–160.
- Tsiatis, A. A., & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88, 447–458.
- Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, 14, 809–834.
- Tsonaka, R., Verbeke, G., & Lesaffre, E. (2009). A semi-parametric shared parameter model to handle non-monotone nonignorable missingness. *Biometrics*, 65, 81–87.
- Vonesh, E. F., Greene, T., & Schluchter, M. D. (2006). Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in Medicine*, 25, 143–163.
- Wu, M. C., & Carroll, R. J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44, 175–188.
- Wulfsohn, M. S., & Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53, 330–339.
- Xu, J., & Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics*, 50, 375–387.
- Yang, W., Israni, R. K., Brunelli, S. M., Joffe, M. M., Fishbane, S., & Feldman, H. I. (2007). Hemoglobin variability and mortality in ESRD. *Journal of the American Society of Nephrology*, 18, 3164–3170.
- Ye, W., Lin, X., & Taylor, J. M. G. (2008). Semiparametric modeling of longitudinal measurements and time-to-event data – a two-stage regression calibration approach. *Biometrics*, 64, 1238–1246.
- Yu, M., Law, N. J., Taylor, J. M. G., & Sandler, H. M. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, 14, 835–862.

This page intentionally left blank

PART V

e-Science methods

This page intentionally left blank

DATA, INFORMATION, KNOWLEDGE, WISDOM

Nancy Staggers and Ramona Nelson

Clinical nursing informatics began in the 1970s with the introduction of computers on clinical units in acute care settings. Today it has expanded to all health care settings and involves social media, mHealth (mobile health), telehealth, ubiquitous sensor data, e-patients, patient portals, and personal health records. With this evolution, attention has shifted from a focus on data entry to data and information extraction, including methods for obtaining meaning from multiple stores of patient data. Today, the data-information-knowledge-wisdom (DIKW) continuum of informatics offers a new perspective for advancing nursing science.

In this chapter, the data, information, knowledge, wisdom (DIKW) continuum is used as a guide for exploring research questions and methods arising from the rapid growth of health information technology (HealthIT). Informatics sources are then outlined and future research directions are proposed.

The data to wisdom model

The concepts of data, information, and knowledge are in common use across the information sciences (Ackoff, 1989; Blum, 1986; Clarke, 1999). In nursing, Graves and Corcoran (1989) first defined nursing informatics as the use of technology and the study of nursing-relevant data, information, and knowledge. This perspective was supplanted by adding the concept of wisdom and considering the flux across the DIKW continuum (Nelson, 2010; Nelson & Joos, 1989). The model depicting these concepts and their interrelationship continues to evolve. The latest depiction of these relationships is shown in Figure 18.1 (Nelson & Staggers, 2014).

The word “data” is central to science, and the words “information,” “knowledge,” and “wisdom” are familiar in everyday language but these words have precise technical definitions in informatics. *Data* refers to symbols, signs, or measures that can be directly captured by a person or machine (Clarke, 1999). This is similar to, but not exactly, the way the term is used in classical science, where “data” is defined as “pieces of information collected in a study” (Polit & Hungler, 1999, p. 699) and they come suffused with meaning as part of the design of a study. *Information* is “data + meaning” and is context-dependent (Clarke, 1999). *Knowledge* is the formalization of relationships, experiences, and rules used to generate information from data (Blum, 1986). *Wisdom* is the appropriate use of knowledge in making decisions to manage or solve human problems (Ackoff, 1989), including nursing problems (Nelson & Joos, 1989). As shown

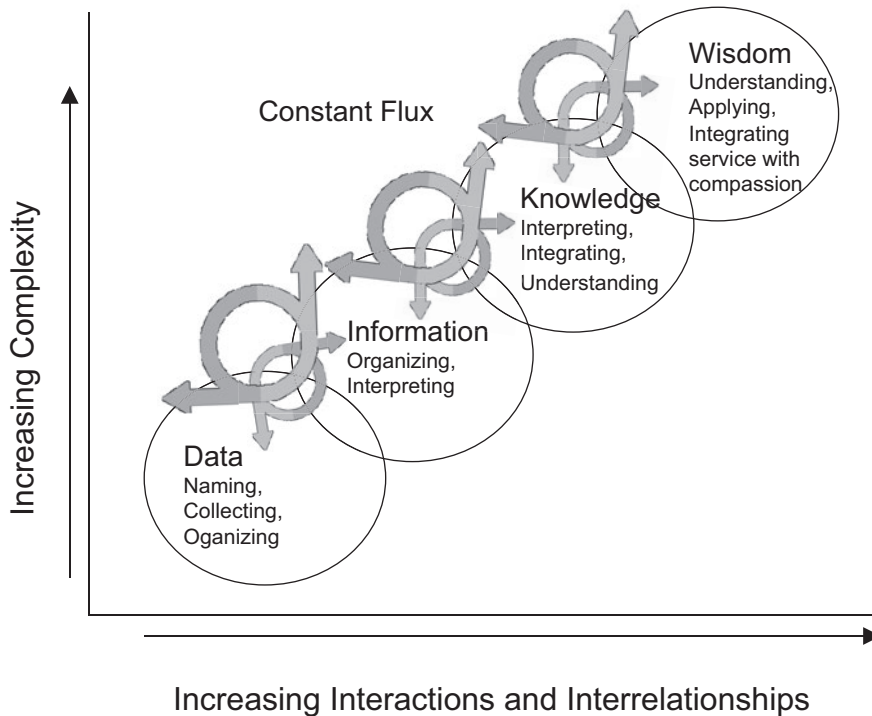


Figure 18.1 The Nelson data to wisdom model. © Ramona Nelson, Ramona Nelson Consulting. Used with permission, all rights reserved.

in Figure 18.1, wisdom is built upon data, information, and knowledge. Movement across the DIKW continuum involves increasing complexity arising from multiple interrelationships and interactions. Analysis of the attributes of the DIKW concepts shows how they overlap and reveals research-related implications.

Data

Symbols, signs, or measures are uninterpreted at the time of capture by persons or machines. Thus, data are meaningless until placed in a context. For example, the number “102” is a symbol that in various contexts may represent a temperature, weight, blood sugar, systolic or diastolic blood pressure, or patient ID code.

In the research process, the research question is defined and data elements of interest are gathered related to the question. In this context data can be conceptualized as the raw material that will be used to answer the research question. However, when considering the role of data within research studies, three major challenges exist. The first involves determining specific data elements to collect. When studying a phenomenon, an event can never be captured in its totality. Rather, data elements are selected by the researcher based on the research question and the research protocol developed by the researcher. However, as with digital photography, the fewer data elements that are collected, the fuzzier the picture of the phenomenon of interest. Following this same analogy a failure to select the right data elements can result in holes in the picture that in turn can lead to erroneous conclusions.

The second challenge concerns data attributes. Data are descriptive or measurable. In the research process the collection of data involves the measurement and/or description of data elements related to phenomenon being studied. These attributes suggest that data are collected via the senses (Frické, 2009). However, the data collection processes of describing and measuring may introduce various forms of error. Devices, including the human senses, are used to derive data and vary in quality; there is no perfect device. Repeated observations are sometimes taken in an attempt to “average out” errors from any one reading. In addition, certain phenomena cannot be directly observed but only inferred from related observations.

The third challenge relates to data quality. The more control the researcher exerts over the data source and the collection process, the more likely the data will be clean. However, researchers walk a tightrope in making these decisions. The more control exerted, the less the data reflect the phenomenon of interest as it exists in reality. In turn, the less likely it is that conclusions based on these data can be generalized to the population of interest. One example is the use of electronic health record (EHR) systems for research. EHRs include large databases capturing care events as they occur and are documented in practice. However, they also often contain human data errors created by busy clinicians.

Information

Information is data that are processed to create meaning, i.e., with context. For example, weight gain over time can be interpreted as normal for a growing child, but abnormal in an aging adult. It is the context and subsequent interpretation that changes a group of numbers from data to information. In another example, a list of heart rates alone would be meaningless, but if these data were tagged with time to create trends, information would emerge. As data are processed, new data items and information can be created. For instance, an average heart rate can become a data element based on several individual measurements or aggregated data elements.

The process by which data are interpreted to create meaning adds an element of judgment and bias. This reality is demonstrated when one reviews the attributes of information. These attributes include: quality (accuracy, coherent, comprehensive, objective, and verifiable), usability (relevant, economical, understandable), format (quantifiable, precise, organized for use), and availability (accessible, secure, timely) (Nelson & Staggers, 2014). If data or information are not available in the qualitative or measureable format needed or in a timely manner, they will not be usable.

Concomitantly, information attributes can also be used to answer four basic questions of importance to all researchers: Is the information of good quality? Is the information usable? Is the required information available? Is the format of the information appropriate?

Knowledge

Knowledge results when the relationships between the data and information are formalized, rather than being just a collection of facts. A knowledge base builds upon the interrelationships between data and information. For example, a textbook organizes a number of facts into a cohesive whole while showing the interrelationships between the facts; thus, it can be seen as containing knowledge.

As Figure 18.1 indicates, the concept of information overlaps with both the concepts of data and knowledge. Research involves both the generation of information from data and the use of information in the creation of new knowledge.

Formalization of the relationships, experiences, and rules used to generate information from data arises from many sources and can be structured in a variety of ways, be related to different topics, and used in different ways. Understanding the broad concept of knowledge begins with

understanding types of knowledge and how different types of knowledge are used. For example, knowledge may be classified as procedural or conceptual. Procedural knowledge involves “knowing how” and is used for actions, such as performing a procedure. Conceptual knowledge involves “knowing about” and is used for explanations.

Graves and Corcoran (1989) identified four types of knowledge used in nursing: empirical, ethical, personal, and aesthetic (also see Carper, 1978). Mantzoukas and Jasper (2008) used a secondary qualitative analysis to identify five types of knowledge used to guide care of hospitalized patients: (a) personal practice knowledge used in therapeutic relationships; (b) theoretical knowledge about relevant facts; (c) procedure knowledge with ready-made answers used effortlessly with activities of providing nursing care; (d) unit or ward cultural knowledge comprising written and unwritten norms, rules, and values for care; and (e) reflexive knowledge that is context-specific and developed through synthesis of the other four types of knowledge through experience.

The larger the knowledge base used to interpret data, the greater amount of material generated. For example, the more knowledge available about normal and abnormal heart rates in different age groups and circumstances, the more information available within a series of heart rate data. Likewise, lack of knowledge can decrease the information within data. For example, a nursing student may view a patient’s liver function tests results and realize these results are abnormal. But an experienced staff nurse looking at the same results will see the information and knowledge, and, in addition, the implications for the patient’s plan of care.

Newly created information can be used to enrich the knowledge base by increasing both the depth and scope of information and knowledge. In a recursive process, then the increased knowledge leads to increased information and potentially new knowledge.

The three attributes of knowledge include accuracy, relevance, and type of knowledge. Lechasseur, Lazure, and Guilbert (2011) used qualitative methods to research the mobilization of knowledge within the critical thinking process of nursing students. They identified nine types of knowledge: intrapersonal, interpersonal, perceptual, moral/ethical, experiential, practical, scientific, contextual, and combinational constructive knowledge. Combinational constructive knowledge was described as “a new type of knowledge, created from a simultaneous combination of all types of knowledge in function of a specific situation” (Lechasseur et al., 2011, p. 1936). Different researchers report higher levels of thinking and practice based on a synthesis of different knowledge types. It is at this point of synthesis that knowledge begins to overlap with wisdom.

Wisdom

Within the DIKW model, wisdom is defined as the appropriate use of knowledge in managing or solving human problems. Wisdom is knowing when and how to use knowledge in managing a need or problem, whether for a patient or other area of interest to nurses. Nelson, (2002) using wisdom effectively in managing a patient problem requires a synthesis of the different types of knowledge and experience while applying these to a variety of real life situations. The introduction of the concept of wisdom gained professional acceptance in 2008 when the American Nurses Association (ANA) included this concept in the Standards of Practice for Nursing Informatics (ANA, 2008). However, little research related to nursing wisdom has been reported.

Information: Sources, issues, and implications

The DIKW model can assist researchers in evaluating data and information sources and evaluating the quality of materials available in health IT products. While a wide variety of data and information sources exist today, EHRs are used as an exemplar. EHRs are a logical source for

clinical research data because they have robust longitudinal data in many settings. Research issues are outlined for this one health IT product, but the same concepts can be applied to a range of health IT applications from patient portals, telehealth, mobile health, or even to administrative information systems.

Applying DIKW to health IT products

Combined within a structure, facts in context allow information to emerge (Tuomi, 1999). Information can be derived computationally by manipulating data using procedures in an organized or structured way. Alternately, information may be generated by marrying data with narrative context.

The notion of information is a critical concept for researchers using health IT products. For example, consider a researcher interested in using clinical information in interdisciplinary discharge summaries to improve the completeness of the summaries. With EHRs, data can be relatively easily extracted to determine current discharge summary content. However, if this material is then analyzed without regard to context (i.e., if it is treated as data rather than information), the product will result in a comprehensive list that has little meaning to clinician users. Discharge summaries and content need to be tailored – at least to the context of specific disciplines, content of interest to all providers and diagnoses, as well as other factors. What is needed in this type of research is *information*; therefore, discharge summary content needs to be extracted within context to create meaning (i.e., information).

EHRs can contain years of longitudinal data about patients, and millions of patient records can be available within and across sites. Before researchers embrace EHRs as a data source, they must understand the potential peculiarities of data integrity and data availability in a specific EHR. This section discusses those topics and also outlines a newer method to help alleviate problems of data integrity and availability for practice-based evidence.

Data/information integrity

The quality of data and information in EHRs can vary substantially. For example, EHRs can have template-based charting and allow copy-forward features. Copy-forward is the ability to copy a previous section of the EHR and edit it – storing it as a new document. This feature is very convenient for users, but can result in errors if careful attention is not paid to editing before the new document is stored. Otherwise, the same data can inadvertently be copied forward for days, creating inaccurate information. For example, the results of the same physical exam were carried forward using this feature for nine days in a local EHR. Not only does this result in an inaccurate representation of the patient's condition, but it is also a vexing problem for researchers due to the difficulty of detecting this kind of error. As with paper charts, missing data can occur in electronic systems, especially if templates and structured documentation are not deployed or if “charting by exception” methods are used. In these instances, normal fields would not be stored in the EHRs database, and an incomplete dataset would result. Last, institutions may have incentives to up-code in order to maximize charges, resulting in data quality issues. As an example, EHR data entry can vary in units of measurement, such as using both pounds and kilograms to record weight on the same pediatric patient. Any trends with these data, such as pediatric growth chart, are then difficult to interpret. Consistency in these data must be rectified to ensure accuracy in interpretation.

Researchers are cautioned to be familiar with EHR functions and understand the characteristics of data and information available from EHRs. Some issues may be detected by asking

users and informaticists about any unintended errors, typical usage patterns, and work-arounds that might cause data integrity issues. In some cases, researchers may want to actually interact with the EHR features that generate the data/information of interest to uncover any potential issues. For example, the current design and usability of electronic medication administration records (eMARs) could contribute to new medication errors (Guo, Irdbarren, Kapsandoy, Perri, & Staggers, 2011). If researchers are interested in evaluating medication errors, especially before and after EHR deployment, understanding the workflow, typical tasks, and eMAR design and its contribution to medication errors would be critical.

Data/information availability

Although EHRs may initially appear to be comprehensive in their data and information about patients, they may not include key data or information of interest to nursing. For example, EHRs do not typically include the quality of life measures of interest to oncology nurse researchers. A study (Guo et al., 2014) about inclusion criteria for participants in randomized control trials (RCTs) for oncology nursing research found that current data standards better support medicine than nursing. This is important because, in an IRB-approved protocol, RCT inclusion criteria could be used as search terms in EHRs to identify potential research participants. Guo et al. stated that inclusion criteria and EHR data elements need to be expanded to consist of at least these factors of interest to nursing: (a) mental health; (b) physical capability; (c) training or education; and (d) occupation.

Other information may be missing in EHRs, such as clinical decision-making methods, sensitive information, or designs to facilitate the overall care for a patient with a particular condition. For example, at shift changes, nurses may give rationale why a particular decision or intervention was completed. The rationale may not be documented in an EHR. Likewise, clinicians may convey sensitive information only verbally, e.g., challenging family dynamics, cancer staging for a patient who did not want to know a prognosis. More important, the nature of most EHRs is that they represent care rendered only in the facilities supported by a specific EHR, while most patients receive care from multiple institutions and providers. Fragmented records of care and limited documentation of follow up for specific problems can occur (Hoffman & Podgurski, 2012). Even within an integrated delivery system, any one EHR is unlikely to represent the whole of care for a patient. Thus, researching patient conditions across settings and their different EHRs is a significant challenge.

Information may also not be collated within EHRs to provide cognitive support for clinicians (Staggers, Clark, Blaz, & Kapsandoy, 2012), and this can also affect the quality of data and information available for research. For example, technology-supported clinical summaries or handoffs may not be available, or they may have inadequate support especially across unit types (Staggers, Clark, Blaz, & Kapsandoy, 2011). Moreover, transitions in care between units and facilities can be just as problematic in the electronic world as they were with paper-based methods.

Practice-based evidence

One method to improve information quality for research using health IT products is a newer technique called *practice-based evidence* (PBE). Unlike evidence-based practice, PBE is a prospective research design using data/information gathered from current practice and stored in health IT products, such as EHRs (Cummins, Pepper, & Horn, 2013). Using this method to construct observational cohort studies, researchers can identify which care processes work in actual clinical settings. PBE studies are distinct from traditional studies in that extensive attention is given to

patient characteristics. Large and realistic samples increase representativeness and external validity. Standard and structured documentation of interventions result from standardized training and quality control checks. Front-line clinicians and patients can be involved in the design, selection of data elements, execution, and analysis for the study.

By using health IT to capture the extensive data across sites, researchers can detect best practices using bivariate and multivariate associations among patient characteristics, process steps, and outcomes. The advantage of PBE study designs is that they capture the complexity of patient and treatment differences in actual settings, and do not alter treatments as would be done in a RCT (Horn & Gassaway, 2007; Horn & Gassaway, 2010). Instead, rigorous attention is paid to standardizing documentation, typically in health IT products, allowing subsequent data analyses to detect associations between treatments and outcomes, while controlling for patient differences. Steps in a PBE study are: (a) create a multisite, multidisciplinary team; (b) control for differences in case mix; (c) implement intensive data collection and check reliability of data; (d) create a study database; (e) test hypotheses successively; and (f) validate and implement study findings. More detailed information about this process is available elsewhere (Cummins et al., 2013; Horn & Gassaway, 2007, 2010).

Knowledge: issues and implications for research

Methods of knowledge generation are discussed next. These include newer literature synthesis techniques, use of data warehouses in health care, and research on nurses as knowledge workers.

Literature synthesis methods

Systematic reviews identify, evaluate, and synthesize all the empirical evidence (Cullum & Dumville, this volume) adhering to prespecified eligibility criteria for a given research question (Cochrane Library, 2013; Higgins & Green, 2011). Specific methods, such as rigorous searches of literature databases, careful analysis and synthesis of studies, are then used to minimize bias in interpreting studies to assure more reliable findings are generated.

Systematic reviews primarily focus on quantitative findings, while nursing research often involves mixed methods (Doyle, this volume) or qualitative studies. While the Cochrane Library does offer methods to synthesize qualitative studies, an integrative review technique may be used. Integrative review is a specific technique for summarizing literature to provide a comprehensive view of a research topic (Whittemore & Knafl, 2005). Unlike systematic reviews, integrative review techniques can encompass all study designs, while still using rigorous processes for searching, analyzing, and synthesizing methods and findings across studies. Methods for integrative reviews are less explicit than systematic reviews; however, guides are available (e.g., Whittemore & Knafl, 2005). Both types of reviews rely on comprehensive searches for pertinent studies. Both methods can also assist in understanding the state of the science for a topic, contribute to theory development, and generate thoughts for practice and policy. The integrative review can be especially useful for newer or emerging phenomena of interest, such as informatics topics – computerized provider order entry issues, mHealth, or challenges with electronic documentation for multidisciplinary teams.

Data warehouses and virtual databases in health care

Data warehouses (DWs) are large repositories of local or regional health information. DWs integrate data across clinical, administrative, departmental, and/or financial information systems. Large collections of health related data may exist in a variety of storage formats that can be termed

a clinical data repository, data repository, data mart, knowledge warehouse, or information warehouse (Lyman, Scully, & Harrison, 2008; Westra, Monsen, & Delaney, this volume). These large databases have the potential to provide extensive benefits for research and quality improvement, as well as clinical and business decision support (Schubart & Einbinder, 2000). They can allow secondary use of the bulk of operational data from a health facility for research. The particular advantage of a data warehouse is that data are longitudinal (linked to specific patients across care episodes and encounters), and can provide population-based views (Lyman et al., 2008). For example, in an early study, researchers at Duke University sampled 45,000 patients in their DW to determine risk factors for preterm birth (Prather et al., 1997). More recently, Botsis et al. (2010) described a DW at Columbia University with 2.7 million records collected since 1994.

Technological advances such as virtual databases have become available. These include cloud computing, which allows virtual linkages across databases. The same concepts from DIKW apply to either DW or virtual databases.

These repositories are increasingly available, although their development in healthcare is fairly recent with the advent of low-cost data storage and increasing computing power. As EHRs are implemented, secondary use of data in DWs and in virtual environments are becoming even more common (Botsis et al., 2010). DWs and virtual structures are not part of the day-to-day operational systems but instead support complex queries that return extensive datasets (Lyman et al., 2008) off-line without negatively impacting care production systems. As a database of databases, the underlying DW data structures (data models and databases) are termed “online analytic processing” and are fundamentally different from operational systems such as EHRs. These structures are designed specifically to optimize complex queries across populations, in contrast to EHR structures that maximize large numbers of small transactions (Lyman et al., 2008).

Some special considerations exist. Smaller facilities may lack the fiscal resources to construct and maintain a DW or virtual databases. Even larger facilities may lack tools and technology to adequately manage large stores of data because very large datasets require highly specialized techniques to store, manage, extract, and analyze. More important, data quality may be inadequate. Substantial preprocessing may be required before analysis (Cios & Nguyen, this volume; Cummins et al., 2013) or even making them unusable.

Data codes and structures

The history of EHRs can affect researcher ability to locate needed information in DWs. Early EHRs were developed to support billing functions so database structures and codes were developed to support those functions. Development of clinical documentation and provider orders functions occurred later. Thus, newer codes to support clinical functions were added to financially focused or ancillary (such as laboratory) structures. Since the initial purpose of some EHRs was to maximize capture of charge data, the database structure may not be designed for optimal retrieval of clinical data. Researchers can benefit from understanding the data model for the system to facilitate the research process.

Standard terms may not exist to facilitate data or information extraction. Without standard terms, researchers must perform multiple searches and/or use multiple search terms to retrieve comprehensive data. For instance, when searching clinical documentation for mention of myocardial infarction, researchers may need to search for “myocardial infarction,” “MI,” “heart attack,” “R/O MI,” as well as symptoms without a concept label, such as “chest pain.” Multiple ICD codes may be needed as well. Structured taxonomies terms, on the other hand, and use of documentation created with structured terms, can allow faster, more comprehensive retrieval because multiple terms are mapped to one term, such as “myocardial infarction.”

Depending upon the topic of interest, researchers will want to determine whether taxonomies are used in their health IT products and their inclusiveness or terms, as this will determine methods and even the feasibility of a study. For example, mobile devices do not currently have standardized terms.

Another limitation is that no standard structure or architecture is available for DWs and virtual databases. Technology architects, if present, at separate facilities are free to design these tools based on their own facility's needs. For researchers using multiple EHRs or DWs, local variations could limit the type of information available across sites, and also impact study timelines.

Secondary data use

Multiple techniques for data extraction are available. Researchers will want to collaborate with clinical analysts or informaticists to create effective searches. Knowledge discovery/data mining (KDDM) is a process of machine learning combined with statistical methods to analyze large amounts of data. A general framework for the KDDM process is outlined in Cios and Nguyen (this volume) and in Cummins et al. (2013).

Research considerations

Data quality, the use of atheoretical research methods with a recursive process of data analysis, and ethical issues need to be considered when generating knowledge using DWs and virtual databases. Data quality issues in EHRs were described earlier. Any poor quality data in source health IT products will be imported into DWs with the potential to exacerbate data quality issues. Thus, data should be assessed for incompleteness, inconsistency, and inaccuracy (Botsis et al., 2010).

Analysis of big data from DWs and virtual databases may fundamentally change research processes of the future. Much of nursing research is currently hypothesis-driven and based on theoretical models. Decisions on what data to collect and the methodology used to collect these data are driven by the hypothesis under study. While some researchers use research questions and theoretically driven models for knowledge discovery, the possibility exists to identify patterns in large datasets instead. This approach also offers an alternative path to knowledge, in that knowledge discovery in databases (KDD) can identify previously unknown systematic data patterns (relationships) across multitudes of patient records, leading to new knowledge.

This inductively based research knowledge discovery process may be disconcerting to some researchers. Some may consider these techniques atheoretical in nature. With this research approach, the researcher does not use a theory, propose a hypothesis, or even ask a research question. Rather, the research question emerges from the data. KDD can be used to uncover unexpected patterns or clusters, allowing information or knowledge to emerge from a large pool of data. A researcher notes the pattern and develops an idea (maybe even a guess) about why the pattern might exist. This idea can lead to additional data collection(s) and/or analyses. Thus, the process from data to idea and back to data to information and knowledge is recursive.

Secondary use of the large datasets is creating new ethical issues. Local DWs are typically governed by local institutional review boards. However, as databases are merged through health information exchanges and virtual methods new issues emerge. For example, Blewett, Parente, Finch, and Peterson (2004) discussed the creation of a national data warehouse to link public and private data to monitor health care trends and costs. Even with de-identified data, how do patients consent to these new purposes of data use, a standard aspect of protocols for protection of the rights of human subjects in research? Since IRBs are local, not national, how does the approval process occur for cross-site data? Currently, it is a researcher burden to apply to each institution,

but with DWs and other large amalgams of information, the new approaches to the IRB process will likely be developed.

Nurses as knowledge workers

A widely accepted premise is that nurses are knowledge workers (Antrobus, 1997; Sorrells-Jones & Weaver, 1999). Knowledge work is nonroutine, unpredictable, nonlinear, multidisciplinary, and involves nonrepetitive activities in a long time frame with changing goals, and the process may be evolving concurrently with product development (Beyerlein, Johnson, & Beyerlein, 1995). This definition is particularly apt for health IT design because informatics research and health IT products often focus on information rather than knowledge levels of support.

A case in point is handoffs or change of shift reports. Recent and past definitions for handoffs emphasize information transfer (Arora et al., 2009; Cohen & Hilligoss, 2010; Stroppe & Ottani, 2006), and do not yet include the concept of knowledge. Moreover, e-handoff designs, such as one from a worldwide vendor of health IT products, may not support nursing handoffs well (Staggers et al., 2011). Analysis of 93 patient handoff transcripts showed the preponderance of this exchange is at the information level (e.g., “No allergies”; “His hematocrit was 40”; and “Next med for you is due at 1900”). Yet, over 40% of the handoff content was classified as knowledge (Matney, Maddox, & Staggers, 2014). How informatics can support a knowledge-based handoff is a subject for future research.

More generally, a future direction for research is how health IT products can be better designed to support knowledge for nurses and teams in activities where information synthesis occurs. Topics include point-of-care issues concerning patient safety (Clarke & Schubert, this volume). For example, how can care transitions be supported through technology? Does verbal repetition ensure the receiving nurse encodes that information? What content would be contained in handoffs if EHRs better supported that activity? What team elements, such as interdisciplinary goals, should be conveyed in health IT products to support the knowledge team of clinical workers?

Wisdom: Issues for research

Although the word “wisdom” appears in a number of nursing publications, a limited number of papers and even fewer research studies are focused on wisdom in nursing. Initial publications discussed wisdom as a separate concept unrelated to data, information, and knowledge. In a seminal paper about “clinical wisdom,” Benner et al. (1999) referred to clinical wisdom as thinking-in-action. Thinking-in-action was defined as “patterns and actions of thought and action that are directly tied to responding to patients and families and the demands of a changing situation and for noticing when clinical assumptions and expectations are not met” (p. 558). Clinical wisdom has also been linked to proficient practice is based on responsibility, thinking and ethical discernment, and a drive for action (Uhrenfeldt & Hall, 2007). In contrast, the term “practical wisdom” in education was proposed to have arisen from encountering, making sense of, and moving through, situations that challenge educators conceptually, culturally, and judgmentally (Paton, 2007).

One of the first nursing publications to use the data-to-wisdom continuum (Nelson, 2002) described a model of care delivery in telephone-based nursing practice. In the third and final phase of the call, wisdom was used to guide the call completion (Schleyer & Beaudry, 2009). Matney and colleagues used the model to point out the different philosophical approaches used in analyzing the concepts of data, information, and knowledge versus the concept of wisdom

(Matney, Brewster, Sward, Cloyes, & Staggers, 2011). Gee and colleagues (2012) used it for understanding how e-patients and clinicians may achieve collective wisdom (Gee et al., 2012). These novel studies suggest that wisdom is a fertile area for additional research. Ideas for additional directions suggested by a review of over 150 papers about wisdom was published in the psychology literature (Standinger & Glück, 2011) and include subjective research about what constitutes wisdom, conceptual models for wisdom, ontogenesis of wisdom, and ways to influence development of wisdom.

Future directions and conclusions

DIKW framework provides future directions in three key areas. First, how can this framework be used in organizing the research process? Could the concepts of knowledge or wisdom be used to present the outcomes and implications for nursing research studies? Second, what research-related questions should be addressed in terms of this framework? What are important attributes for wisdom? Third, how will the tools of technology change the research process? For example, will e-patients on social networking sites become collaborators with nurses in creating the concept of collective knowledge or wisdom?

The data-to-wisdom continuum is an underutilized framework in nursing research. Researchers have expanding sources of data and information but these are not without challenges involving data integrity and availability. The reader is invited to develop DIKW-based questions within key areas: *Is there evidence of knowledge and wisdom in nursing practice? What are their manifestations? How can health IT products be better designed to support the knowledge and wisdom of nurses as knowledge workers?*

References

- Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16, 3–9.
- American Nurses Association. (2008). *Nursing informatics: Scope and standards of practice*. Washington, DC: Author.
- Antrobus, S. (1997). Developing the nurse as a knowledge worker in health – learning the artistry of practice. *Journal of Advanced Nursing*, 25, 829–835.
- Arora, V. M., Manjarrez, E., Dressler, D. D., Basaviah, P., Halasyamani, L., & Kripalani, S. (2009). Hospitalist handoffs: A systematic review and task force recommendations. *Journal of Hospital Medicine*, 4, 433–440.
- Benner, P., Hooper-Kyriakidis, P., & Stannard, D. (1999). *Clinical wisdom and interventions in critical care: A thinking-in-action approach*. Philadelphia, PA: Saunders.
- Beyerlein, M. M., Johnson, D. A., & Beyerlein, S. T. (Eds.). (1995). *Knowledge work in teams (Advances in interdisciplinary studies of work teams)* (Vol. 2). Greenwich, CT: Jai Press.
- Blewett, L. A., Parente, S. T., Finch, M. D., & Peterson, E. (2004). National health data warehouse: Issues to consider. *Journal of Healthcare Information Management*, 18, 52–58.
- Blum, B. I. (1986). Clinical information systems – A review. *Western Journal of Medicine*, 145, 791–797.
- Botsis, T., Hartvigsen, G., Chen, F., & Weng, C. (2010). Secondary use of EHRs: Data quality issues and informatics opportunities. *AMIA Summits on Translational Science Proceedings*, 2010, 1–5.
- Carper, B. A. (1978). Fundamental patterns of knowing in nursing. *Advances in Nursing Science*, 1, 13–24.
- Clarke, R. (1999). Fundamentals of “information systems.” Retrieved from www.rogerclarke.com/SOS/ISFundas.html
- Cochrane Library. (2013). What is a systematic review? Chichester: Wiley. Retrieved from www.thecochranelibrary.com/view/0/AboutCochraneSystematicReviews.html
- Cohen, M. D., & Hilligoss, P. B. (2010). The published literature on handoffs in hospitals: Deficiencies identified in an extensive review. *Quality and Safety in Health Care*, 19, 493–497.
- Cummins, M. R., Pepper, G. A., & Horn, S. D. (2013). Knowledge discovery, data mining, and practice-based evidence. In R. Nelson & N. Staggers (Eds.), *Health informatics: An interprofessional approach* (pp. 54–71). St. Louis, MO: Elsevier.

- Frické, M. (2009). The knowledge pyramid: A critique of the DIKW hierarchy. *Journal of Information Science*, 35, 131–142.
- Gee, P. M., Greenwood, D. A., Kim, K. K., Perez, S. L., Staggers, N., & DeVon, H. A. (2012). Exploration of the e-patient phenomenon in nursing informatics. *Nursing Outlook*, 60, e9–e16.
- Graves, J. R., & Corcoran, S. (1989). The study of nursing informatics. *Image: Journal of Nursing Scholarship*, 21, 227–231.
- Guo, J.-W., Irdar, S., Kapsandoy, S., Perri, S., & Staggers, N., (2011). eMAR user interfaces: A call for ubiquitous usability evaluations and product redesign. *Applied Clinical Informatics*, 2, 202–224.
- Guo, J.-W., Sward, K. A., Beck, S. L., Wong, B., Staggers, N., & Frey, L. (2014). Using a content analysis to identify study eligibility criteria concepts in cancer nursing research. *Computers Informatics Nursing*, 32, 333–342.
- Higgins, J.P.T., & Green, S. (Eds.) (2011). *Cochrane handbook for systematic reviews of interventions*. Version 5.1.0. Chichester: The Cochrane Collaboration. Retrieved from <http://handbook.cochrane.org/>
- Hoffman, S., & Podgurski, A. (2012). Big bad data: Law, public health, and biomedical databases. *Journal of Law, Medicine and Ethics: Case Legal Studies*, Research Paper No. 2012–2034. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2168931
- Horn, S.D., & Gassaway, J. (2007). Practice-based evidence study design for comparative effectiveness research. *Medical Care*, 45, S50–S57.
- Horn, S.D., & Gassaway, J. (2010). Practice based evidence: Incorporating clinical heterogeneity and patient-reported outcomes for comparative effectiveness research. *Medical Care*, 48, S17–S22.
- Lechasseur, K., Lazure, G., & Guilbert, L. (2011). Knowledge mobilized by a critical thinking process deployed by nursing students in practical care situations: A qualitative study. *Journal of Advanced Nursing*, 67, 1930–1940.
- Lyman, J. A., Scully, K., & Harrison, J. H., Jr. (2008). The development of health care data warehouses to support data mining. *Clinics in Laboratory Medicine*, 28, 55–71.
- Mantzoukas, S., & Jasper, M. (2008). Types of nursing knowledge used to guide care of hospitalized patients. *Journal of Advanced Nursing*, 62, 318–326.
- Matney, S., Brewster, P. J., Sward, K. A., Cloyes, K. G., & Staggers, N. (2011). Philosophical approaches to the nursing informatics data-information-knowledge-wisdom framework. *Advances in Nursing Science*, 34, 6–18.
- Matney, S., Maddox, L., & Staggers, N. (2014). Nurses as knowledge workers: Is there evidence of knowledge in patient handoffs? *Western Journal of Nursing Research*, 36(2), 171–190.
- Nelson, R. (2002). Major theories supporting health care informatics. In S. P. Englehardt & R. Nelson (Eds.), *Health care informatics: An interdisciplinary approach* (pp. 3–27). St. Louis, MO: Mosby.
- Nelson, R. (2010). Nursing informatics: Educational implications. In L. Caput (Ed.), *Teaching nursing: The art and science* (pp. 653–693). Glen Ellyn, IL: College of DuPage Press.
- Nelson, R., & Joos, I. (1989, Fall). On language in nursing: From data to wisdom. *PLN Vision*, 6.
- Nelson, R., & Staggers, N. (2014). Theoretical foundations of health informatics. In R. Nelson & N. Staggers (Eds.), *Health informatics. An interprofessional approach* (pp. 18–39). St. Louis, MO: Elsevier Mosby.
- Paton, B. (2007). Knowing within: Practice wisdom of clinical nurse educators. *Journal of Nursing Education*, 46, 488–495.
- Polit, D.F., & Hungler, B.P. (1999). *Nursing research: Principles and methods* (6th ed.). Philadelphia, PA: Lippincott, Williams, & Wilkins.
- Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., & Hammond, W. E. (1997). Medical data mining: Knowledge discovery in a clinical data warehouse. *Proceedings of the AMLA Annual Fall Symposium*, 101–105.
- Schleyer, R., & Beaudry, S. (2009). Data to wisdom: Informatics in telephone triage nursing practice. *AACN Viewpoint*, 31(5), 1–13.
- Schubart, J. R., & Einbinder, J. S. (2000). Evaluation of a data warehouse in an academic health sciences center. *International Journal of Medical Informatics*, 60, 319–333.
- Sorrells-Jones, J., & Weaver, D. (1999). Knowledge workers and knowledge-intensive organizations, Part 1: A promising framework for nursing and healthcare. *Journal of Nursing Administration*, 29, 12–18.
- Staggers, N., Clark, L., Blaz, J. W., Kapsandoy, S., (2011). Why patient summaries in electronic health records do not provide the cognitive support necessary for nurses' handoffs on medical and surgical units: Insights from interviews and observations. *Health Informatics Journal*, 17, 209–223.

- Staggers, N., Clark, L., Blaz, J. W., & Kapsandoy, S. (2012). Nurses' information management and use of electronic tools during acute care handoffs. *Western Journal of Nursing Research*, 34, 153–173.
- Standinger, U. M., & Glück, J. (2011). Psychological wisdom research: Commonalities and differences in a growing field. *Annual Review of Psychology*, 62, 215–241.
- Strople, B., & Ottani, P. (2006). Can technology improve intershift report? What the research reveals. *Journal of Professional Nursing*, 22, 197–204.
- Tuomi, I. (1999). Data is more than knowledge: Implications of the reversed knowledge hierarchy for knowledge management and organizational memory. *Journal of Management Information Systems*, 16, 103–117.
- Uhrenfeldt, L., & Hall, E. O. C. (2007). Clinical wisdom among proficient nurses. *Nursing Ethics*, 14, 387–398.
- Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing*, 52, 546–553.

19

BIG DATA IN NURSING RESEARCH

Bonnie L. Westra, Karen A. Monsen, and Connie W. Delaney

In the US and worldwide, nurses constitute the largest segment of the healthcare workforce, with central roles in leading change and advancing health (Institute of Medicine [IOM], 2010a, 2010b). Likewise, nursing documentation and assessments generated during patient encounters in clinics, general and specialized hospital units, workplaces and schools, homes, the streets – wherever people receive care – constitute a massive data source with critical information about patient/consumer health status, nursing care and interventions, outcomes, safety, patient experiences, access to care, utilization of health services, and costs. This vast assemblage of data becomes a treasure trove when used in the process of knowledge discovery to ensure innovation, quality, safety, and value in health care (IOM, 2011).

Electronic health records (EHRs) are a major source of big data for research. EHRs replace paper charts and go beyond the limitations of paper by streamlining and optimizing workflow processes, prompting clinicians with evidence-based guidelines for decision-making, and providing data for secondary use. EHRs contain current and historical information such as patient histories, test results, care provided, and outcomes documented by all clinicians involved in care. In the US, more than \$19 billion has been invested to rapidly implement and meaningfully use EHRs for patient care, health information exchange across the continuum of care, business analytics, and research (Murphy, 2011).

Standardization of terminologies and a robust informatics infrastructure are not only essential to support the transformation of health care documentation within the EHR but also are essential to supporting health information exchange; accessing comparable data across settings, health systems, and EHRs; and ensuring access to and use of nursing and healthcare data for research. The purpose of this chapter is to (a) describe standardized terminologies that support interoperability and meaningful use of nursing data, including research, especially as incorporated in electronic health records (EHRs); (b) provide examples of big data nursing research using knowledge discovery in databases; (c) describe international initiatives in nursing informatics; and (d) consider use of nursing and inter-professional big data in learning healthcare systems designed to provide the “best care at lower costs” (IOM, 2013).

Standardized terminologies

Nursing data describe patient care and the contexts in which nurses work. The type or focus of data collected by and about nurses varies across practice settings. Data of various types represent patient characteristics, nursing care, interprofessional, and management concerns that are often integrated into EHRs. Understanding the integration process is essential to retrieving and linking nursing and other health care data for research. It also provides insights about how to map data collection protocols onto standardized nursing terminologies.

EHR data may be structured or unstructured. Structured concepts provide consistency in care documentation and also create data that can be aggregated and used in research. Structured data represent concepts (terms; e.g., pain, mobility) that are characterized using a name, definition, and unique code. A variety of concepts, structures, and levels of granularity are used in structured systems; examples illustrated in Figures 19.1 and 19.2 show the Nursing Interventions Classification (Dochterman & Bulechek, 2004) and the Omaha System, respectively. Structured data are more valuable when terms are coded using national standards rather than customized to local codes (Bowles et al., 2013).

Longstanding concern about consistent use of language within and across nursing theory, research, and practice motivated establishment of national standards for coding nursing and other health data (Clark & Lang, 1992; Webber, 2010; Westra, Delaney, Konicek, & Keenan, 2008). Nursing terminologies (classifications, nomenclatures) and minimum data sets (Werley, Devine, Zorn, Ryan, & Westra, 1991) meet international criteria for inclusion in information systems when they: (a) are clinically relevant and support the domain of nursing; (b) have concepts that are unambiguous, with a single unique identifier; and (c) have been tested for reliability, validity, and clinical usefulness (Coenen, McNeil, Bakken, Bickford, & Warren, 2001).

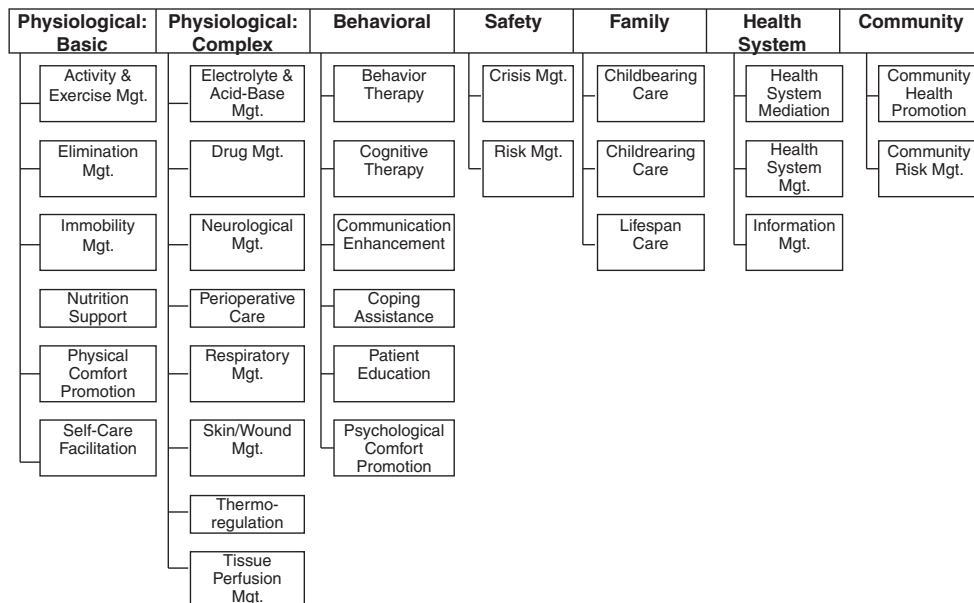


Figure 19.1 A high-level structure for the Nursing Interventions Classification.

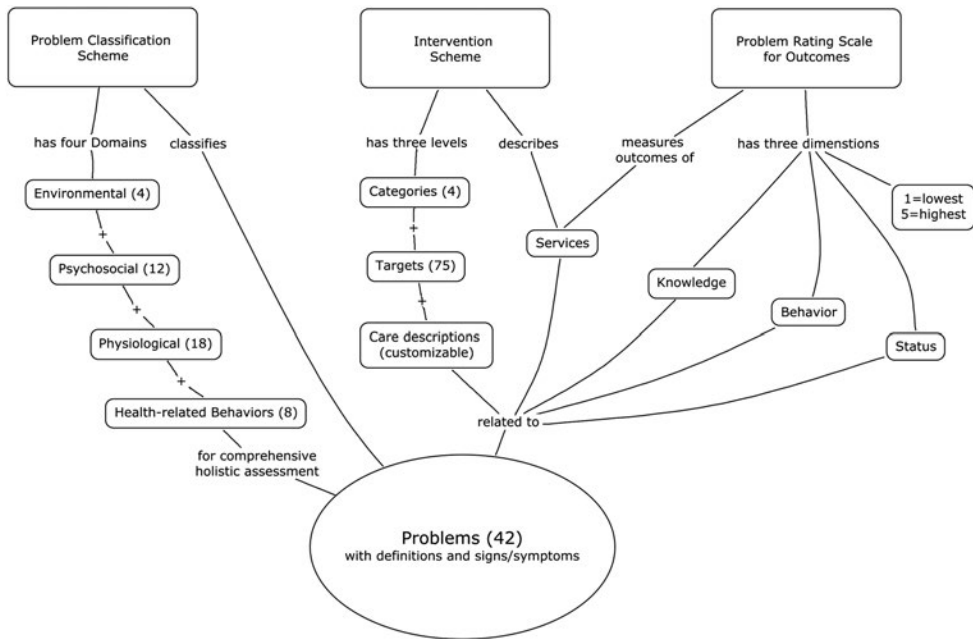


Figure 19.2 Omaha System ontology.

Information about nursing minimum data sets and terminologies that meet international standards is listed in Tables 19.1 and 19.2. Nursing minimum data sets provide high-level frameworks essential to describing the practice and context of care, whereas nursing terminologies are used to document clinical and administrative data at a more granular level. Elements of the Nursing Minimum Data Set (NMDS) characterize nursing care, patient/client characteristics, and services (Werley & Lang, 1988); the Nursing Management Minimum Data Set (NMMDS) describes the care environment and nursing resources (Huber, Schumacher, & Delaney, 1997). As shown in Table 19.2, a uniform terminology to completely represent nursing practice does not exist. Rather, a unified approach across terminologies – the Unified Medical Language System (www.nlm.nih.gov) – is used to link nursing terminologies with interdisciplinary terminologies. Examples include RxNorm (medication codes), LOINC (Logical Observation Identifiers Names and Codes including laboratory data, names for types of documents, clinical assessments, and management data), ICD (International Classification of Diseases including medical diagnoses and procedures), or SNOMED-CT (Systematized Nomenclature of Medicine including clinical terms for medical diagnoses, observations, and findings).

Subsets

Subsets of concepts from a terminology that describe a single NMDS data element, such as nursing diagnoses, can be used to integrate nursing terminologies in EHRs. For example, the SNOMED-CT problem list can be used for nursing problems because it is consistent with US federal requirements for meaningful use of EHRs and supports an integrated patient-centered problem list (Matney et al., 2012). Subsets are also used with the International Classification for

Table 19.1 Comparison of elements: nursing minimum data set and nursing management minimum data set

<i>Nursing Minimum Data Set^a</i>	<i>Nursing Management Minimum Data Set^{b,c}</i>
<i>Nursing care</i> • Nursing diagnosis • Nursing intervention • Nursing outcome • Intensity of nursing care <i>Patient/client demographics</i> • Personal identification^d • Date of birth^d • Sex^d • Race and ethnicity^d • Residence^d <i>Services</i> • Unique identifier: facility/service agency^d • Unique identifier: patient/client • Unique identifier: principal RN provider • Date: encounter or episode admission^d • Date: discharge or termination^d • Disposition of patient/client^d • Expected payer for most of this bill^d	<i>Environment</i> • Unit/service unique identifier • Type of nursing delivery unit/service • Patient/client population • Volume of nursing delivery unit/service • Care delivery structure and outcomes • Patient/client accessibility • Nursing delivery unit/service accreditation <i>Nursing Resources</i> • Staffing • Satisfaction • Nurse demographics per unit or service • Clinical mental work • Environmental condition • EHR implementation stage <i>Financial Resources^e</i> • Payer type • Reimbursement • Budget: nursing unit/service • Expenses

Note: EHR = electronic health record; RN = registered nurse.

^aWerley & Lang (1988).

^bHuber, Schumacher, & Delaney (1997).

^cThe Nursing Management Minimum Data Set (NMMDS) was updated with completion in 2015 and submission to LOINC for standardized coding and dissemination. Many variables changed and definitions, subelements and codes were updated. The complete listing can be found at www.nursing.umn.edu/icnp/center-projects/minimum-data-sets/index.htm. Data element changes were: (a) NMMDS 7 Clinical Decision Making Complexity was replaced by NMMDS 20 Clinical Mental Work; (b) NMMDS 8 Environmental Complexity was replaced by NMMDS 21 Environmental Condition and NMMDS 22 Electronic Health Record (EHR) Implementation Stages; (c) NMMDS 9 – Autonomy was discontinued; (d) NMMDS 11 Management Demographic replaced by NMMDS 19; (e) NMMDS 12 Staff Demographic Profile replaced by NMMDS 19 Nurse Demographics per Unit or Service; and (f) NMMDS Financial Domain Variables were not included in updates as these are not available at the unit level.

^dElements comparable to those in the Uniform Hospital Discharge Data Set (UHDDS).

Nursing Practice (ICNP), which combines nursing diagnoses with interventions and outcomes to represent the care of specific populations such as children with HIV/AIDS in developing countries (Choromanski, Collins, Hart, Westra, & Delaney, 2012) or palliative care for dignified dying in community centers (International Council of Nurses [ICN], n.d.). Creation of subsets makes it easier to use larger terminologies in a consistent manner.

Clinical models (templates, clinical content models, archetypes, detailed clinical models, domain analysis models) use small sets of related concepts for consistent documentation. Detailed clinical models (DCMs) comprise entity-attribute-value triplets that provide reusable standardized data. Entities (like pain) associated with attributes (like severity or frequency) have values (ranging

Table 19.2 Nursing minimum data sets, terminologies, and billing codes

<i>Type</i>		<i>Example/Website</i>
Minimum data set	NMDS	Nursing Minimum Data Set http://z.umn.edu/datasets
	NMMDS	Nursing Management Minimum Data Set http://z.umn.edu/datasets USANMMDS/home.html
Clinical terminology	CCC	Clinical Care Classification www.sabacare.com/
	NANDA-I	North American Nursing Diagnoses Association International ^a www.nanda.org
	NIC	Nursing Interventions Classification www.nursing.uiowa.edu/excellence/nursing_knowledge/clinical_effectiveness/index.htm
	NOC	Nursing Outcomes Classification www.nursing.uiowa.edu/excellence/nursing_knowledge/clinical_effectiveness/index.htm
	Omaha System	Omaha System ^{a,b} www.omahasystem.org
	PNDS	Perioperative Nursing Data Set www.aorn.org
	ICNP	International Classification of Nursing Practice www.icn.ch/pillarsprograms/international-classification-for-nursing-practice-icnpr/
Reference terminology	LOINC	Logical Observation Identifiers Names and Codes http://loinc.org
	SNOMED-CT	Systematic Nomenclature of Medicine Clinical Terms www.ihtsdo.org/snomed-ct/
Billing	ABC	Alternative Billing Codes www.abccodes.com

^aNursing diagnoses.^bMultidisciplinary ontology.

from *never* to *continuously*, for example) (Goossen, Goossen-Baremans, & van der Zel, 2010). In Korea, DCMs are coded using the ICNP, whereas in the US, SNOMED CT or LOINC are frequently used. Examples are DCMs for perinatal care (Park, Min, Jeon, & Chung, 2012) and obesity management (Kim & Park, 2011).

Use of EHR data for research requires standardization of content and documentation processes. When nursing data are not coded in EHRs using national standards, researchers must map nursing data recorded with local codes to standardized data formats. The Unified Medical Language System (UMLS) is useful for mapping terms from evidence-based practice (EBP) care plans to one or many national data standards (Demner-Fushman et al., 2008; National Library of Medicine [NLM], 2013). Most nursing terminologies and data sets can be accessed through the UMLS. The UMLS integrates terminologies, coding standards, and related resources to

promote dissemination, use, and interoperability through a machine-readable vocabulary knowledge source, the UMLS Metathesaurus (<https://uts.nlm.nih.gov/home.html>). A license for the UMLS Metathesaurus can be obtained at no charge. Licensees have access to the UMLS Terminology Service applications (including a Metathesaurus Browser, Semantic Network Browser, and SNOMED CT Browser) and may download data files including UMLS Knowledge Sources, RxNorm updates, and CORE Problem List and Route of Administration Subsets of SNOMED CT. Some material in the UMLS Metathesaurus is copyright protected. Users of the UMLS Metathesaurus are solely responsible for compliance with any copyright, patent, or trademark restrictions and are referred to the copyright, patent, or trademark notices appearing in the original sources.

Terminologies and clinical guidelines are used to represent EBP within an EHR. The Clinical Practice Model (CPM) framework in acute care (www.cpmrc.com) and Zynx Health resources for ambulatory care (www.zynxhealth.com/) are examples of how guidelines for clinician reference, integration of guidelines into flowsheets or order sets, and alerts for clinical decision support can undergird evidence-based care and generate data for unique patients using standardized formats. The Omaha System is also structured using EBP care plans for community-based practices; EBP care plans are defined as EBP guidelines linked with nursing diagnoses or focus of care and interventions for a specific clinical condition or process of care (e.g., pain management, discharge planning) to achieve outcomes (Monsen, Foster et al., 2011; also see <http://omahasystemguidelines.org>). EBP care plans are clinical decision support tools that guide nursing practice from planning care through documentation and data retrieval for research

Knowledge discovery in nursing databases

Approach

Knowledge discovery in databases (KDD) is a data-driven, computationally intensive e-science enterprise (Cios & Nguyen, this volume; Hey, Tansley, & Tolle, 2009; Staggers & Nelson, this volume). The goal is discovery of new knowledge using large, complex, and highly dimensional data sets. KDD in nursing is enabled by ready availability of EHR data, emerging computational processing, and computer technologies and architectures. Problems in nursing science are complex and multidimensional. Aggregated information from clinical databases using nursing minimum data sets and terminologies is the foundation for knowledge discovery in nursing databases and is essential for understanding what has happened (or will happen) in sickness and in health.

Data mining (Cios & Nguyen, this volume; Fayyad, Grinstein, & Wierse, 2001; Westra, Dey et al., 2011) and visualization techniques (Fayyad, Grinstein, & Wierse, 2001; Kopanitsa, 2012) are two approaches to discovering knowledge from databases. They support discovery of new knowledge when analyzing the large, complex nursing and health data sets needed to tailor (personalize) nursing care and are essential to making genomic discoveries (Conley, this volume) used to personalize medications and other medical treatments. Data mining combines methods from statistics, machine learning, and computer science with clinical expertise to discover new knowledge that is well suited to the high dimensionality of complex intervention data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Berger & Berger, 2004). Visualization is a graphical display of data that enables pattern detection and communication (Fayyad et al., 2001). Both data mining and visualization enable researchers to identify novel hypotheses from patterns that would not have been discovered using traditional research methods.

Big data sources

Clinical data repositories (CDRs), also called data warehouses, provide centralized storage of clinical and other data for outcome analysis, quality improvement, and research. Variables from the NMDS and NMMDS integrated with other health professionals and patient data stored in CDRs can be retrieved and used to discover new knowledge. Developed in 1994, the National Database for Nursing Quality Indicators (NDNQI) integrated comparable nursing data across US hospitals to support patient safety and quality improvement (American Nurses Association, 2013; Montalvo, 2007). Data are available to researchers. (See www.nursingquality.org/). Examples of research based on NDNQI data include: pressure ulcers (Bergquist-Beringer, Gajewski, Dunton, & Klaus, 2011), patient falls (Lake, Shang, Klaus, & Dunton, 2010), and nurse turnover (Staggs & Dunton, 2012).

Funding through US Clinical Translational Science Awards (CTSAs) for intense health research centers develop CDRs to support research; however, the type of data included varies across CTSAs. The content of CDR data increasingly includes EHRs as well as administrative data, clinical trial data, registries, social security or death index, and genomic or proteomic data (MacKenzie, Wyatt, Schuff, Tenenbaum, & Anderson, 2012). The multidimensional design; extraction (data retrieval from a source), transfer, and loading (placing data in a CDR) process; transformation (mapping values from a storage system to those usable in a database query); and materialization strategy (creating a table of results from a database query) for inclusion of nursing diagnoses, interventions, and outcomes in CDRs are critical to enabling analytical processing (Hylock, Street, Lu, & Currim, 2008). At this time, however, nursing data are largely missing in CTSA CDRs. Exceptions are the Universities of Minnesota and Kansas, which are collaborating to code care plan and flowsheet data for inclusion in their CTSA CDRs using SNOMED-CT and LOINC codes (Waitman, Warren, Manos, & Connolly, 2011; Westra, Delaney, Melton-Meaux, & Monsen, 2012). This effort will likely expand to additional CTSAs as new methods and tools are shared.

Software needed to implement standardized nursing terminologies in practice is sold by commercial vendors (Table 19.2). Once the software is implemented, databases are populated with clinical input about patient status and care provided. Vendors may be contacted to obtain assistance in recruitment of their agency customers as collaborators in research (e.g., Dierich, Mueller, & Westra, 2011). Another approach used to recruit 781 homecare agencies was to distribute information about the study using emails and flyers through professional organizations, conferences, software vendors, and journal advertising (Bliss, Westra, Savik, & Hou, 2013).

Privacy and consent

Secondary use of EHR and other protected health information is governed in the United States by the Health Insurance Portability and Accountability Act (HIPAA) of 1996's Privacy Regulation (www.hhs.gov/ocr/hipaa) and the Common Rule (Nosowsky & Giordano, 2006; Bova, Drexler, & Sullivan-Bolyai, 2012). Patients have the right to privacy of their information (PHI); therefore use of patient data in a CDR requires approval of a HIPAA Privacy Board and/or institutional review boards (IRB). Patients have the right to opt out of being contacted for research recruitment or use of their data for research when it includes personally identifiable information. States and institutions may choose to have more stringent laws for use of patient data for research and either require patients to explicitly opt in or opt out to use their data for research. Therefore, patient consent in an EHR needs to be tracked and data queries should include verifying consent before providing PHI data to researchers. With approval of the IRB or Privacy

Board with oversight of the research, patient data may be used without an individual's authorization, if the data are either de-identified or a limited data set such that there is minimal risk of re-identification or harm to patients. Laws, regulations, and guidelines on human protections in research vary internationally ("International Compilation," n.d.). Investigators do not have authority to determine whether data usage is subject to regulation and approval, so guidance should always be sought from IRBs or Privacy Boards before beginning a project.

Predictive analytics is shifting research to learning algorithms in real time within EHRs to guide clinicians to identify patients at risk for adverse events such as pressure ulcers or hospital readmissions. The ethical dilemma is that the algorithms provide population-based probabilities applied to an individual to determine whether an intervention may be effective can bias clinician decisions, particularly for patients with multiple chronic health conditions and health disparities. Therefore, a four-phase life cycle for predictive analytics is recommended for reducing bias in algorithms and creating transparency for decision making (Cohen, Amarasingham, Shah, Xie, & Lo, 2014). Recommendations address issues with acquiring data, such as use of EHR data that is provided without patient consent (but has IRB approval) to prevent bias. A second recommendation is that when building and validating models, patients should be involved in governing boards to determine which problems are of highest priority from their perspective. Additionally, rigorous standards for validation are needed. Creating transparency in the variables used as well as testing the models across institutions increases the generalizability of predictive models. Close monitoring of the impact of any predictive analytics is essential for assuring that intended outcomes are achieved and unintended consequences are minimized or prevented. Finally, broader dissemination is needed with equitable access to predictive analytics.

Exemplars

The Center for Nursing Informatics at the University of Minnesota leads discovery, application, and cutting-edge thinking for nursing and health informatics scholarship to improve the health of individuals and communities (www.nursing.umn.edu/icnp/index.htm). The purpose of this Center is to (a) discover and employ innovative methods of informatics research in nursing; (b) use standardized nursing terminologies and essential minimum data sets for knowledge discovery; and (c) apply research methods to clinical and other information systems. The Center serves as the home for the NMDS (Werley & Lang, 1988), the NMMDS (Huber, Schumacher, & Delaney, 1997); the Omaha System Partnership (Martin, 2005; Omaha System Partnership, 2015); and the ICN ICNP Research and Development Center (a global collaborative of researchers).

Intervention patterns among complex homecare patients

Studying intervention patterns among complex homecare patients is one example of big data nursing science. Clinical data documented in the Omaha System was utilized to discover nursing intervention clusters used during homecare (Monsen, Westra, Yu, Ramadoss, & Kerr, 2009). The data involved 621,386 interventions provided to 2,862 complex homecare patients. The purpose was to identify hidden patterns (clusters) of interventions used during homecare. The analysis was completed using unsupervised learning (see Cios & Nguyen, this volume) with *k*-means clustering and estimation maximization (Witten & Frank, 2005, pp. 254–265) in the Weka 3.5 environment. (See www.cs.waikato.ac.nz/~ml/weka/ for more information about Weka.) During subject matter expert review of analytic results, 88 of 150 clusters were reduced to 24 groups composed of 144 interventions reflecting (a) assisting patients to stay at home safely

[low intensity], (b) basic care for a health condition [low intensity], and (c) comprehensive care for a health condition [high intensity].

Client and intervention patterns in home visiting data

The problems, interventions, and outcomes classification schemes in the Omaha System were used to discover client and intervention patterns during home visits for 484 high-risk mothers (Monsen, Banerjee, & Das, 2010). Data obtained on at least three home visits for each mother were used. Following preprocessing, clients were clustered using mixed membership naïve Bayes methods (Shan & Banerjee, 2008), and 51,256 temporally defined interventions were clustered using KMETIS, a graph partitioning method (Karypis & Kumar, 1998). Twenty empirically defined intervention clusters were reduced to 14 meaningful clusters by clinical experts. All clusters included Teaching, Guidance and Counseling and Surveillance interventions; case management interventions were included in all but two intervention clusters. Evidence that intervention clusters were differentially associated with client clusters emerged, suggesting that nurses tailored interventions to family need.

Visualizing interventions over time

Temporal patterns in nursing interventions for 1,204 adolescent and adult female clients visited by 15 nurses were visualized (Kim, Monsen, & Pieczkiewicz, 2013) using D3js streamgraphs (Bostock, Ogievetsky, & Heer, 2011). Visual perceptual review resulted in 29 different temporal intervention patterns that appeared to vary by problem severity, complications, and number of home visits; post hoc standard statistical analysis showed that type of problem defined nearly half of the patterns; and both severity ($p < .001$), and number of visits ($p < .001$) were significantly associated with a pattern shape.

International initiatives

Standardization and international use of terminologies facilitate global efforts in comparative effectiveness research and knowledge discovery in nursing. Most nursing terminologies have been translated into multiple languages for global application. Many nursing terminologies are hosted by international organizations (e.g., NANDA, the Omaha System, the International Health Terminology Standards Development Organization that owns and maintains SNOMED CT, and the International Council of Nursing that owns ICNP).

International nursing minimum data sets (i-NMDS)

i-NMDSs that describe the health status of national populations and care provided by nurses are increasingly used for research (Goossen et al., 2006). There is no single i-NMDS, rather, there are recommended essential data elements for country adoption. The International Council of Nursing and the International Medical Informatics Association-Nursing Informatics Special Interest Group agreed to co-sponsor the development of i-NMDS (Goossen et al., 2006). Four countries (USA, Belgium, France, and Australia) have developed i-NMDS and others (Finland, the Netherlands) have tested a NMDS in the general hospital setting. A comparison across international NMDSs shows similarities in nursing care data item as well as descriptors for patients, nurses, and institutions; however, there are also differences in data elements as well as level of granularity. The majority have a shared purpose of describing and comparing nursing

care as well as support of research and management. Individual countries need to be contacted for access to these data.

Omaha System

The Omaha System (Martin, 2005) was developed in the US and adopted in community practice settings like homecare, public health, and school health. Use expanded nationally and internationally to diverse settings (faith-based nursing, hospital-based and managed-care case management, acute care and rehabilitation hospitals, occupational health, and long-term care). Omaha System user groups promote and support practice quality, documentation quality, information management, program evaluation, and knowledge discovery (Topaz, Golfenshtein, & Bowles, 2014). The Omaha System Partnership includes multidisciplinary scientific teams with experience in data analytics; affiliate members from many countries who contribute clinical data, suggest important clinical questions, and work with the scientific team on research and evaluation projects; and a warehouse of de-identified clinical Omaha System (Omaha System Partnership, 2015).

International Classification for Nursing Practice (ICNP)

The ICNP, recognized by the World Health Organization, arose from extensive international effort since 1989 based on terms nurses across the world used to describe their practice (Clark & Lang, 1992; ICN, n.d.) Updated every two years, the terminology is available without charge in languages including Arabic, Brazil-Portuguese, English, Farsi (Persian), French, German, Icelandic, Indonesian, Italian, Japanese, Korean, Mandarin, Norwegian, Polish, Portuguese, Rumanian, Slovak, Spanish, and Swedish.

ICNP is used for research (www.icn.ch/). Brazilian investigators evaluated the usefulness of the ICNP to describe nursing diagnoses for patients with infectious diseases (de Andrade, da Nóbrega, Freire, & Nóbrega, 2013) and heart failure (Lins, Santo, Fuly, P., & Garcia, 2013). Korean investigators demonstrated interoperability of ICNP terms to SNOMED CT (So & Park, 2011), fit of the ICNP for perinatal care (Park, Min, Jeon, Kim, & Km, 2012), and methods of implementing the ICNP in an EHR (Park, Min, Jeon, & Chung, 2012; Park, Min, Jeon, Kim, & Km, 2012). Publications related to ICNP can be found at www.icn.ch/images/stories/documents/programs/icnp/icnp_references.pdf

Learning health systems

Continuously learning health systems use digital capture of care experiences to guide, support, tailor, and improve health services (IOM, 2010a). Designing and adopting a learning health-care system involves methodically generating and applying the best evidence for the collaborative healthcare choices of each patient and provider; drives the process of discovery as a natural outgrowth of patient care; and ensures innovation, quality, safety, and value in health care (IOM, 2013). Nursing's preparedness for these bold expectations is paramount; audacious nursing engagement and leadership are essential to this commitment to transforming healthcare and generating practice-based evidence.

Clinical data and a robust digital infrastructure are staples of learning health systems and the foundations for continuously learning health systems (IOM, 2010a, 2011). Developing a patient-centric learning healthcare system builds on years of nursing discovery arising from standardization of nursing terminologies; identification of essential nursing data; and development of accessible, comparable nursing big data (Moen & Knudsen, 2013). Technology including health information technology is a key component of innovations in the US national nursing

Table 19.3 Nursing informatics agenda for 2008–2018^a

Nursing Informatics Priorities

- Build upon knowledge gained in nursing concept representation to address genomic and environmental data
 - Guide the reengineering of nursing practice
 - Harness new technologies to empower patients and their caregivers for collaborative knowledge development
 - Develop user configurable software to support complex data visualization, analysis, and predictive modeling
 - Facilitate the development of middle-range nursing informatics theories
 - Encourage innovative evaluation methodologies attentive to human-computer interface factors and organizational context
 - Include interdisciplinary researchers
-

^aBakken, Stone, & Larson (2008).

research strategic plan (National Institute of Nursing Research [NINR], 2011); a nursing informatics agenda has been articulated (Table 19.3; Bakken, Stone, & Larson, 2008); both will further propel nursing's capacity to contribute to and lead development of a continuously learning health system to optimize care.

The central focus on the patient/consumer is foundational to learning health systems and nursing leadership and engagement in learning health systems. The patient-centric value is exemplified by initiatives such as PatientsLikeMe, launched in 2004, which currently engages more than 200,000 patients addressing 1,800 diseases (www.patientslikeme.com/). PatientsLikeMe, a strong advocate for unlocking patients' own data, is building an open platform for anyone who wants to develop new outcome measures with patients through funding from the Robert Wood Johnson Foundation. Those measures will be made freely available. All aspects of the initiative are congruent with a patient-centric, learning health system.

Conclusion

Research using big data from EHRs is rapidly evolving. New perspectives and methods for knowledge discovery are emerging for analyzing big data. Development of standardized terminologies and minimum data sets over the past 40 years positions nursing for advancement of big data research. Resources for mapping existing data to standardized nursing terminologies for research across health systems and software vendors are available and use of subsets of terminologies for more consistent data collection facilitate research use has already resulted in novel findings not possible using standard scientific approaches. Transformation in healthcare, a knowledge discovery infrastructure that drives quality and safety, and a patient-centric focus combine to support bold nursing engagement and action through the evolution of nursing science, nursing informatics, and interprofessional engagement across research, education, and clinical practice.

References

- American Nurses Association. (2013). NDNQI. Retrieved from www.nursingquality.org/
- Bakken, S., Stone, P. W., & Larson, E. L. (2008). A nursing informatics research agenda for 2008–18: Contextual influences and key components. *Nursing Outlook*, 56, 206–214.
- Berger, A. M., & Berger, C. R. (2004). Data mining as a tool for research and knowledge development in nursing. *CIN: Computers, Informatics, & Nursing*, 22, 123–131.

- Bergquist-Beringer, S., Gajewski, B., Dunton, N., & Klaus, S. (2011). The reliability of the National Database of Nursing Quality Indicators pressure ulcer indicator: A triangulation approach. *Journal of Nursing Care Quality*, 26, 292–301.
- Bliss, D. Z., Westra, B. L., Savik, K., & Hou, Y. (2013). Effectiveness of wound, ostomy and continence-certified nurses on individual patient outcomes in home health care. *Journal of Wound, Ostomy, and Continence Nursing*, 40, 135–142.
- Bostock, M., Ogievetsky, V., & Heer, J. (2011). D(3): Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17, 2301–2309.
- Bova, C., Drexler, D., & Sullivan-Bolyai, S. (2012). Reframing the influence of the Health Insurance Portability and Accountability Act on research. *Chest*, 141, 782–786.
- Bowles, K. H., Potashnik, S., Ratcliffe, S. J., Rosenberg, M., Shih, N. W., Topaz, M., . . . Naylor, M. D. (2013). Conducting research using the electronic health record across multi-hospital systems: Semantic harmonization implications for administrators. *Journal of Nursing Administration*, 43, 355–360.
- Choromanski, L., Collins, B. J., Hart, C. M., Westra, B., & Delaney, C. (2012). Creating an ICNP subset: Children with HIV/AIDS in developing countries. *CIN: Computers, Informatics, Nursing*, 30, 183–189.
- Clark, J., & Lang, N. (1992). Nursing's next advance: An internal [international] classification for nursing practice. *International Nursing Review*, 39(4), 109–112.
- Coenen, A., McNeil, B., Bakken, S., Bickford, C., & Warren, J. J. for the American Nurses Association Committee on Nursing Practice Information Infrastructure. (2001). Toward comparable nursing data: American Nurses Association criteria for data sets, classification systems, and nomenclatures. *Computers in Nursing*, 19, 240–246.
- Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs*, 33, 1139–1147.
- de Andrade, L. L., da Nóbrega, M. M., Freire, M. E., & Nóbrega, R. V. (2013). Diagnósticos de enfermagem para clientes hospitalizados em uma clínica de doenças infectocontagiosas [Nursing diagnoses for clients hospitalized in an infectious disease clinic]. *Revista Da Escola De Enfermagem Da U S P*, 47, 448–455.
- Demner-Fushman, D., Seckman, C., Fisher, C., Hauser, S. E., Clayton, J., & Thoma, G. R. (2008). A prototype system to support evidence-based practice. *AMIA Annual Symposium Proceedings, 2008*, 151–155.
- Dierich, M. T., Mueller, C., & Westra, B. L. (2011). Medication regimens in older home care patients. *Journal of Gerontological Nursing*, 37(12), 45–55.
- Dochterman, J. M., & Bulechek, G. M. (Eds.). (2004). *Nursing interventions classification (NIC)* (4th ed.). St. Louis, MO: Mosby.
- Fayyad, U., Grinstein, G. G., & Wierse, A. (2001). *Information visualization in data mining and knowledge discovery*. San Francisco, CA: Morgan Kaufmann.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37–54.
- Goossen, W. T., Delaney, C. S., Coenen, A., Saba, V. K., Sermeus, W., & Warren, J. J. . . . Casey, A. (2006). The International Nursing Minimum Data Set (i-NMDS). In C. A. Weaver, C. W. Delaney, P. Weber, & R. L. Carr (Eds.), *Nursing and informatics for the 21st century: An international look at cases, practice, and the future* (pp. 305–320). Chicago, IL: HIMSS.
- Goossen, W. T., Goossen-Baremans, A., & van der Zel, M. (2010). Detailed clinical models: A review. *Health-care Informatics Research*, 16, 201–214.
- Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm. Data-intensive scientific discovery*. Redmond, WA: Microsoft Research.
- Huber, D., Schumacher, L., & Delaney, C. (1997). Nursing Management Minimum Data Set (NMDS). *Journal of Nursing Administration*, 27(4), 42–48.
- Hylock, R., Street, W. N., Lu, D.-F., & Currim, F. (2008). NursingCareWare: Warehousing for nursing care research and knowledge discovery. In J. Li, D. Aleman, & R. Sikora (Eds.). *Proceedings of the 3rd INFORMS Workshop on Data Mining and Health Informatics*. Washington, DC: INFORMS.
- Institute of Medicine. (2010a). *Clinical data as the basic staple of health learning: Creating and protecting a public good: Workshop summary*. Washington, DC: National Academies Press.
- Institute of Medicine. (2010b). *The future of nursing: Leading change, advancing health*. Washington, DC: National Academies Press.
- Institute of Medicine. (2011). *Digital infrastructure for the learning health system: The foundation for continuous improvement in health and health care: Workshop Series Summary*. Washington, DC: National Academies Press.

- Institute of Medicine. (2013). *Best care at lower cost: The path to continuously learning health care in America*. Washington, DC: National Academies Press.
- International Compilation of Human Research Standards. (n.d.). Washington, DC: US Department of Health and Human Services. Retrieved from www.hhs.gov/ohrp/international/intlcompilation/intlcompilation.html
- International Council of Nurses. (n.d.). ICN programs: eHealth Programme. Retrieved from www.icn.ch/images/stories/documents/pillars/Practice/icnp/eHealth_-_ICNP.pdf
- Karypis, G., & Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20, 359–392.
- Kim, E., Monsen, K. A., Pieczkiewicz, D. S. (2013). *Visualization of Omaha System data enables data-driven analysis of outcomes*. Poster presented at the Annual Meeting of the American Medical Informatics Association, Washington, DC.
- Kim, Y., & Park, H. A. (2011). Development and validation of detailed clinical models for nursing problems in perinatal care. *Applied Clinical Informatics*, 2, 225–239.
- Kopanitsa, G. (2012). Standard based multiclient medical data visualization. *Studies in Health Technology and Informatics*, 180, 199–203.
- Lake, E. T., Shang, J., Klaus, S., & Dunton, N. E. (2010). Patient falls: Association with hospital Magnet status and nursing unit staffing. *Research in Nursing & Health*, 33, 413–425.
- Lins, S. M., Santo, F. H., Fuly P., & Garcia, T. R. (2013). Subconjunto de conceitos diagnosticos da CIPE(R) para portadores de doenca renal cronica [Subset of ICNP(R) diagnostic concepts for patients with chronic kidney disease]. *Revista Brasileira de Enfermagem*, 66, 180–189.
- MacKenzie, S. L., Wyatt, M. C., Schuff, R., Tenenbaum, J. D., & Anderson, N. (2012). Practices and perspectives on building integrated data repositories: Results from a 2010 CTSA survey. *Journal of the American Medical Informatics Association*, 19(e1), e119–e124.
- Martin, K. S. (2005). *The Omaha System: A key to practice, documentation, and information management* (Reprinted 2nd ed.). Omaha, NE: Health Connections Press.
- Matney, S. A., Warren, J. J., Evans, J. L., Kim, T. Y., Coenen, A., & Auld, V. A. (2012). Development of the nursing problem list subset of SNOMED CT®. *Journal of Biomedical Informatics*, 45, 683–688.
- Moen, A., & Knudsen, L. M. M. (2013). Nursing informatics: Decades of contribution to health informatics. *Healthcare Informatics Research*, 19(2), 86–92.
- Monsen, K. A., Banerjee, A., & Das, P. (2010). Discovering client and intervention patterns in home visiting data. *Western Journal of Nursing Research*, 32, 1031–1054.
- Monsen, K. A., Foster, D. J., Gomez, T., Poulsen, J. K., Mast, J., Westra, B. L., & Fishman, E. (2011). Evidence-based standardized care plans for use internationally to improve home care practice and population health. *Applied Clinical Informatics*, 2, 373–383.
- Monsen, K. A., Westra, B. L., Yu, F., Ramadoss, V. K., & Kerr, M. J. (2009). Data management for intervention effectiveness research: Comparing deductive and inductive approaches. *Research in Nursing and Health*, 32, 647–656.
- Montalvo, I. (2007). The National Database of Nursing Quality Indicators® (NDNQI®). *Online Journal of Issues in Nursing*, 12(3), Manuscript 2.
- Murphy, J. (2011). HITECH programs supporting the journey to meaningful use of EHRs. *CIN: Computer, Informatics, and Nursing*, 29, 130–131.
- National Institute of Nursing Research. (2011). *Bringing science to life. NINR strategic plan*. NIH publication #11–7783. Bethesda, MD: Author. Retrieved from www.ninr.nih.gov/aboutninr/ninr-mission-and-strategic-plan#.VICTwzGjNcY
- National Library of Medicine (NLM). (2013). Unified Medical Language System® (UMLS®). Retrieved from www.nlm.nih.gov/research/umls/
- Nosowsky, R., & Giordano, T. J. (2006). The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule: Implications for clinical research, *Annual Review of Medicine*, 57, 575–590.
- Omaha System Partnership. (2015). Completed research. Retrieved from <http://omahasystempartnership.org/completed-research/>
- Park, H.-A., Min, Y. H., Jeon, E., & Chung, E. (2012). Integration of evidence into a detailed clinical model-based electronic nursing record system. *Healthcare Informatics Research*, 18, 136–144.
- Park, H.-A., Min, Y. H., Jeon, E., Kim, Y., & Kim, H. Y. (2012). Development of an electronic nursing records system based on information models and clinical practice guidelines. *Studies in Health Technology and Informatics*, 180, 1206–1208.

- Shan, H., & Banerjee, A. (2008). Mixed-membership naïve Bayes models. Technical report TR 09-002. Unpublished paper. Minneapolis: University of Minnesota, Department of Computer Science and Engineering.
- So, E.-Y., & Park, H.-A. (2011). Exploring the possibility of information sharing between the medical and nursing domains by mapping medical records to SNOMED CT and ICNP. *Healthcare Informatics Research*, 17, 156–161.
- Staggs, V.S., & Dunton, N. (2012). Hospital and unit characteristics associated with nursing turnover include skill mix but not staffing level: An observational cross-sectional study. *International Journal of Nursing Studies*, 49, 1138–1145.
- Topaz, M., Golfenshtein, N., & Bowles, K. H. (2014). The Omaha System: A systematic review of the recent literature. *Journal of the American Medical Informatics Association*, 21, 163–170.
- Waitman, L. R., Warren, J. J., Manos, E. L., & Connolly, D. W. (2011). Expressing observations from electronic medical record flowsheets in an i2b2 based clinical data repository to support research and quality improvement. *AMIA Annual Symposium Proceedings, 2011*, 1454–1463.
- Webber, P.B. (2010). Language consistency: A missing link in theory, research and reasoning? *Journal of Advanced Nursing*, 66, 218–227.
- Werley, H. H., Devine, E. C., Zorn, C. R., Ryan, P., & Westra, B. L. (1991). The nursing minimum data set: Abstraction tool for standardized, comparable, essential data. *American Journal of Public Health*, 81, 421–426.
- Werley, H. H., & Lang, N. M. (Eds.). (1988). *Identification of the Nursing Minimum Data Set*. New York, NY: Springer.
- Westra, B. L., Delaney, C. W., Konicek, D., & Keenan, G. (2008). Nursing standards to support the electronic health record. *Nursing Outlook*, 56, 258–266.e1.
- Westra, B. L., Delaney, C. W., Melton-Meaux, G., & Monsen, K. A. (2012, March). *Using electronic community care data to enrich translational health research*. Poster presented at the 2012 Summit on Clinical Research Informatics, San Francisco, CA.
- Westra, B. L., Dey, S., Fang, G., Steinbach, M., Kumar, V., Savik, K., . . . Dierich, M. (2011). Data mining techniques for knowledge discovery from electronic health records. *Journal of Healthcare Engineering*, 2(1), 55–74.
- Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann.

DATA MINING AND DATA VISUALIZATION

Krzysztof J. Cios and Dat T. Nguyen

The goal of data mining (DM), also called data analytics or predictive analytics, is *to make sense of big data*. To quote from a recent NSF (n.d.) posting, the term *big data* “refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future.” The main challenges in making sense of big data are the number of data points (size), number of features/attributes describing the data (dimensionality of the space), number of values each feature takes on, and heterogeneity of data types (images, signals, text/interviews, health records, etc. (Cios & Moore, 2002; Cios, Pedrycz, Swiniarski, & Kurgan, 2007; Wu, Zhu, Wu, & Ding, 2013). Thus, DM is not about analyzing small, structured data sets that can be modeled using classical techniques (Hand, 1998; Henly, this volume).

The goal of this chapter is to introduce data mining and data visualization in the context of the knowledge discovery process (KDP) described later, where approaches to data preprocessing methods, model building, and model validation are identified. Techniques for visualizing high-dimensional data in two- or three-dimensional spaces are also outlined. Last, links to some implementations of the described algorithms are provided for readers interested in using them.

Machine learning is the acquisition of structural descriptions for data from examples (data points) that can then be used for prediction, explanation, and understanding (Witten, Frank, & Hall, 2011). Supervised learning refers to finding mapping between inputs and outputs when both are known, whereas unsupervised learning refers to identification of hidden structures (clusters) when only inputs are known.

Data mining proceeds under various scenarios. The easiest scenario occurs when information about features as well as class membership (category) are known for all data points. This data situation is reflected in the form of data pairs such as (dataaboutpatient1, diseasetype2), (dataaboutpatient2, diseasetype5), etc. Data with this configuration are called training or “supervised” and many existing DM techniques are good at performing supervised learning tasks. A DM algorithm that works well on both small (say up to 100,000 data points) and big data (say over 100 million) is called *scalable*. At the other end of the spectrum are data where only information about patients is available (dataaboutpatient1, dataaboutpatient2, etc.). In this case, class is unknown and unsupervised learning is used. Only a few techniques are available for modeling unsupervised data; key approaches are clustering and association rules. In the middle of the spectrum are data where information (say, about disease diagnosed) exists only for a very small fraction of patients (data points). Techniques appropriate to working with

this type of data are called semi-supervised learning. In this chapter we are concerned only with supervised learning techniques for model building.

Finding a good model of the data, which at the same time is easy to understand, is at the heart of DM. For instance, an easy to understand rule-based model is:

*IF patient has a bed sore
THEN move the patient on her side.*

In this example, the input data (the IF part) are the features (attributes; symptoms) while the output (the THEN part) is an action to be taken.

Knowledge discovery process

Organizations including hospital and health systems and clinical translational science institutes obtain and store voluminous data (Hylock, Street, Lu, & Currim, 2008; MacKenzie, Wyatt, Schuff, Tenenbaum, & Anderson, 2012). Domain experts such as nurse scientists use big data to gain new information and knowledge to solve problems in better ways (such as better ways to treat patients and obtain better health outcomes) (Berger & Berger, 2004; Staggers & Nelson, this volume; Westra, Delaney, & Monsen, this volume). This is what is meant by saying “making sense” of data. Discovering new knowledge from data (the knowledge discovery process) is a highly interactive and iterative process, the success of which depends heavily on access to domain knowledge. Thus, data miners work closely with domain experts in the knowledge discovery process.

Big data are of no value without means for efficient and effective extraction of new information from them. New knowledge extracted from data is usually presented by data miners to data owners (domain specialists) in the form of a model. Ideally, a model is an easy to understand summary of the data that supports use of data for some purpose (such as better treatment protocols of patients leading to faster discharge from a hospital), but not all models are equally easy to understand. Machine learning techniques (understood here ONLY as rule or decision tree algorithms) generate models in the format of simple, English-like IF . . . THEN . . . rules. Others, such as neural network algorithms, may have good predictive power but as the models themselves may be difficult to interpret.

In DM, the modeling process is data driven as opposed to classical statistical situations where modeling is typically theory driven. In DM, data size is large and the goal is to build a data model that is small (not too complex) but that still describes the data well whereas in classical statistics, researchers design studies using the smallest sample size that gives sufficiently confident estimates of model parameters (Hand, 1998).

Standardization of the KDP has evolved from a nine-step process (Fayyad et al., 1996) through the business-focused cross-industry standard process for data mining (CRISP-DM; Shearer, 2000) to the six-step process shown in Figure 20.1 (Cios & Kurgan, 2005). Each step of the KDP, after the first, can start only after successful completion of the previous step and usually has many feedback loops.

Steps

1. Understanding the problem domain

This step involves working with domain experts to understand and define the problem, and to decide on goals for solving the problem. It also includes learning about currently used solutions. For a data miner, one of the barriers to overcome is to learn domain-specific terminology, which may not be trivial. Importantly, project goals need to be translated into data-mining goals.

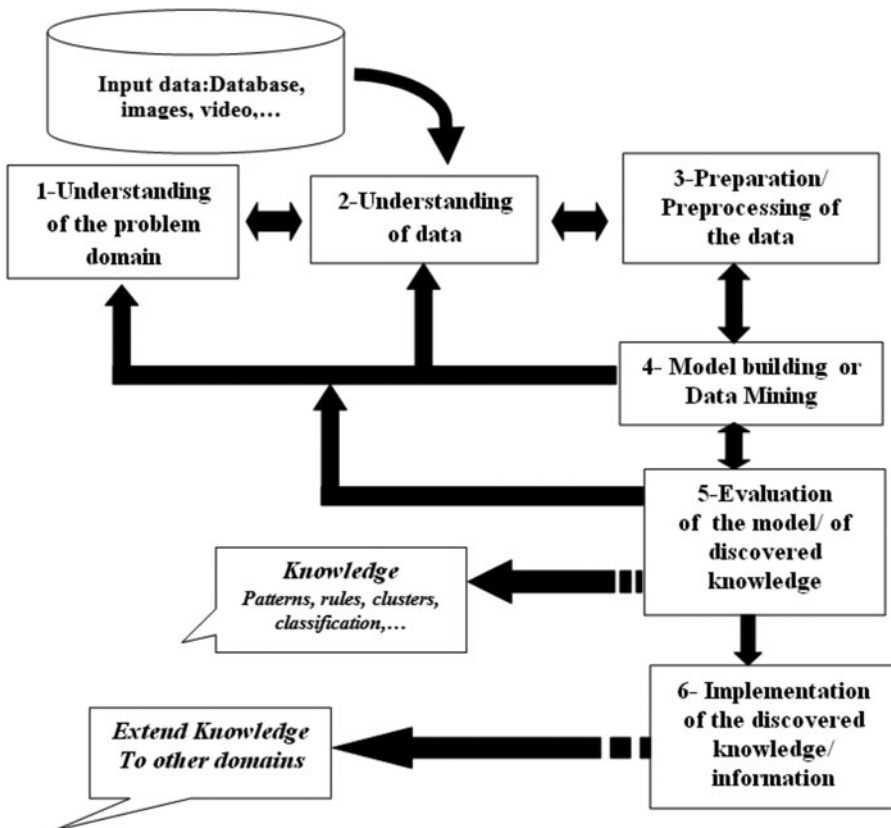


Figure 20.1 The six-step knowledge discovery process.

For instance, a nurse may say he is interested in cases where there is a significant improvement in controlling blood glucose, reflected in hemoglobin A1c levels in people with diabetes. For a data miner it would mean looking at features characterizing people with diabetes and deciding what “significant” means (is it any decrease in A1c, or only a decrease to a value less than 7%?).

2. Understanding the data

This step includes finding out what data are available and obtaining exemplar data. The more background knowledge (like about a disease) a data miner learns, the better. Data are checked for completeness, redundancy, plausibility of attribute values (e.g., valid ranges of hemoglobin A1c levels). Usefulness of the available data for solving the problem identified in Step 1 must be assessed.

For example, information extracted from data captured using the Nursing Management Minimum Data Set (NMMDS; Huber, Schumacher, & Delaney, 1997) can be used to understand nursing resource allocation (staffing) and the impact on implementation of safety measures and patient outcomes (Westra et al., 2010). Data miners working with NMMDS domain experts would need to understand the data elements, definitions, and codes used to document nursing management actions in order to appropriately process and model the data.

3. Preparation/preprocessing of data

This is the most important and time-consuming step. Dirty data (missing, wrong, or nonstandard representations; Kim, Choi, Hong, Kim, & Lee, 2003) are cleaned, operations performed, possibly many times, and associations are studied to decide which data will be used as input for model building in Step 4. Cleaned and selected data can then be processed using feature selection algorithms (choosing a subset from hundreds of original features) or feature extraction algorithms (transforming the original feature space into a new feature space that is not interpretable by humans), discretization, and other techniques.

4. Model building or actual data mining

Many algorithms for building models can be used to reveal new information or knowledge hidden in data. Importantly, new patterns unknown to domain experts may emerge (Hand & Bolton, 2004). For example, a study of preferences of potential patients for various hospital services using cluster analysis showed that the market comprised four segments: “old-fashioneds, value conscious, affluent, and professional want-it-alls” (Woodside, Nielsen, Walters, & Muller, 1988). Neither practicing nurses nor nurse scientists would likely have characterized patients in this way prior to the analysis.

5. Evaluation of the model or the discovered knowledge

Model evaluation or validation has two aspects. One is for a data miner to use formal techniques to assess quality of generated model(s). The other is to present the model (information/knowledge) to the domain expert who always has the final say on whether what was discovered is truly novel and thus useful. Formal evaluation includes calculating measures of model quality before presenting them to the experts. Only the best model approved by domain experts from a set of generated models is retained and used in practice.

6. Implementation of the discovered information/knowledge

The final step in the knowledge discovery process is exclusively in the purview of the domain expert. It involves application of the discovered knowledge in the domain. For example, home care records of 2,072 episodes of care for 1,793 elderly patients were used to predict improvement in urinary and bowel incontinence; a total of 265,966 interventions were documented; among other findings, decision tree analysis showed that interventions for cognitive and emotional functioning, functional status, and physiological variables were associated with improvement in bowel incontinence; and findings were expected to be used to aid clinical decision making based on rules discovered in the study (Westra et al., 2011).

KDP: interactive and iterative

As shown by the feedback loops in Figure 20.1, steps in the KDP are highly iterative and interactive. The loop from *understanding the data* to *understanding the problem domain* allows incorporation of existing domain knowledge. The loop from *preparation of data* to *understanding of data* reflects the use of specific information about the data in selection of appropriate data preprocessing algorithms. The loop from *data mining* to *understanding the problem domain*

allows a response to unsatisfactory DM results, which may suggest the need for restating project goals. *Data mining* loops back to *understanding of data* because poor initial understanding of the data may have resulted in incorrect selection and subsequent failure of a DM method. Going forth and back between *data mining* and *preparation of data* is needed to ensure optimal data preparation for the data mining step. The loop from *evaluation of discovered knowledge* to *understanding the problem domain* allows for domain expert-judged invalidity of the discovered knowledge arising from incorrect understanding or interpretation of the domain, incorrect design or incorrect understanding of problem restrictions, requirements, or goals. When invalidity occurs, the entire knowledge discovery process must be repeated from scratch. The loop from *evaluation of discovered knowledge* to *data mining* is executed when discovered knowledge is not novel. The simplest solution is to choose a different DM method and repeat the DM step.

KDP: time and effort

Estimation of the time and effort needed to complete each step is critical to planning a DM project. Data preparation/preprocessing is generally judged to be the most time-consuming, requiring about 50–60% of the entire effort, followed by the understanding of the data step that requires about 25% of the entire effort (Pal & Jain, 2005; Cios & Kurgan, 2005). Intensive interaction with domain experts is needed throughout the process.

Preprocessing

Preprocessing is an important and very time-consuming step. Some preprocessing methods are general, whereas others are more applicable to specific kinds of data (e.g., to signals or images). We describe only methods for feature selection and discretization.

Feature selection

A *feature* (attribute), or characteristic of an observed unit (data point), is an observed variable or a transformation of a raw observed variable(s) that may be composed of numbers, nominal values, alphanumeric strings, or graphs (Cios et al. 1998; Cios et al., 2007; Guyon & Elisseeff, 2003). Feature selection (FS) is a process of finding the best feature subset X_{opt} using some predefined criterion, from the original set of features. FS is different from feature extraction, which involves transformation of original features into a new space using variance maximization, dimension reduction techniques, such as principal components analysis (Pearson, 1901).

Feature selection is a central data mining activity, particularly in life sciences where it is important to retain original attributes in model building. Feature selection is also needed when original feature set consists of so many features (possibly hundreds or thousands) that examination of all of them is computationally too expensive, or some features are irrelevant (providing no further information), redundant (providing the same information), contain noise, or may have been recorded in error.

Figure 20.2 depicts phases of a typical feature selection process. Class membership must be known to perform feature selection (e.g., diagnosis status: with or without a disease). In the first phase, a subset of features is selected for evaluation. The initial subset may contain no features, all features, or any random subset of features. The selected current subset of features is evaluated

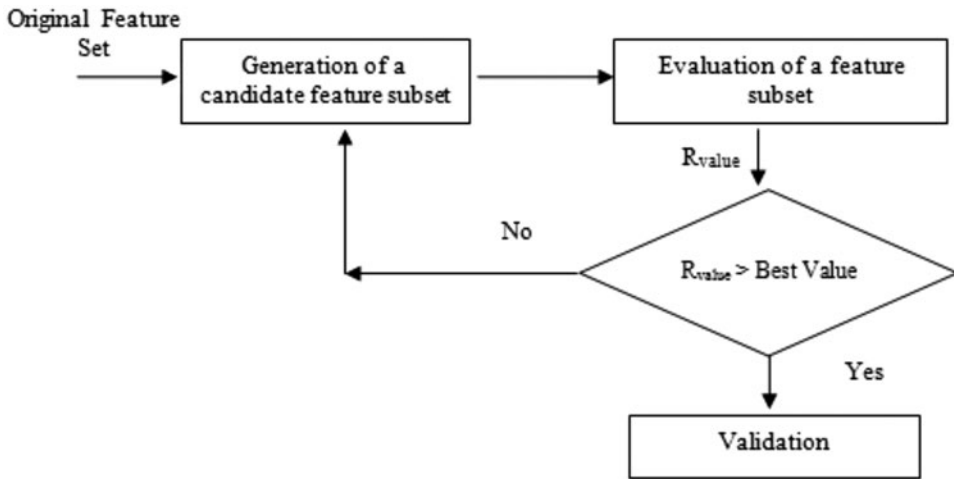


Figure 20.2 Phases of the feature selection process.

Table 20.1 Evaluation functions and feature selection algorithms

Evaluation Function	Exemplar Algorithms	Generality	Time Complexity	Accuracy
Euclidean Distance	Relief Branch and Bound	Yes	Low	N/A
Entropy, Information Gain	Decision Tree Minimal Description Length	Yes	Low	N/A
Correlation Coefficient	Probability of Error Average Correlation Coefficient	Yes	Low	N/A
Min-features Bias	Focus Las Vegas	Yes	Moderate	N/A
Classifier Error	Sequential Forward Selection Sequential Backward Selection	No	High	High

Note: Generality refers to usefulness for classifiers. Time complexity is how long it takes to perform an evaluation. Accuracy is a measure of classification error. N/A means that the accuracy cannot be calculated. From Kudo and Sklansky (2000).

using some feature selection criterion; then, features are added, removed, or replaced with a new subset. The process is repeated until a best set of features is identified.

The evaluation phase of the feature subset is performed, using a classifier, as follows:

$$\begin{aligned}
 R_{\text{value}} &= J(\text{candidate subset}) \\
 \text{if } (R_{\text{value}} > \text{best\_value}) &\text{ then best\_value} = R_{\text{value}}
 \end{aligned}$$

where R_{value} is a value of an evaluation function J and current candidate subset is the argument for J at each iteration. As shown in Table 20.1, many different evaluation functions may be used including Euclidean distance, information (entropy, information gain), dependency (correlation), feature bias, and classifier error. The search for the most optimal feature subset is stopped when a certain value (such as threshold or maximum value) of a criterion function is reached. Classifiers

are used for validation of the selected feature subset by, for instance, looking at the resulting accuracy of classification (diagnosis). Notice that some classifiers can be used for both feature selection and for evaluation of a subset of features for their predictive power.

Let us look at a spirometry data set available from the National Center for Biotechnology Information (NCBI) web site (www.ncbi.nlm.nih.gov/). The spirometry data set includes 182 features like temperature (Celsius), forced vital capacity (FVC) grade, forced expiratory volume (FEV) grade, plateau achieved (yes or no), body temperature-pressure-saturation (BTPS) factor, calibration time, calibration code, age at spirometry, test position, extrapolated volume (ml), forced expiratory flow 25% (ml/sec), forced expiratory flow 50% (ml/sec), and forced expiratory flow 75% (ml/sec). We ran feature selection algorithms (from WEKA; see section on Software) such as best first in subset evaluation (Hall, 1999), relief evaluation (Kira & Rendell, 1992), and greedy stepwise in wrapper subset evaluation (Kohavi & John, 1997) and found 28 significant (most important) features including FVC, FEV, peak expiratory flow (PEF), maximum voluntary ventilation (MVV), slow vital capacity (SVC), and total lung capacity (TLC). This reduction of original features by more than 80% makes subsequent analysis using reduced dimensionality data much easier.

Different feature selection methods have different properties. The Relief algorithm can handle noisy features but not redundant or correlated features, the Focus algorithm can detect redundant features but not noisy features, etc. Remember that a universally best feature selection method does not exist, so finding a good feature subset is a challenging task. A good feature subset should simplify data description, reduce the cost of data collection, and improve accuracy of a classifier.

Discretization

Let us recall that dimensionality refers to three aspects of a data set: the number of data points, the number of features, and the number of values each feature takes on. To address the number of values each feature may take on, one first checks the types of features that describe the data. Often they are nominal, discrete, continuous, or combinations of them. It is advantageous, and often a required preprocessing step for many classifiers, to discretize continuous features. Key discretization methods are described next.

Continuous numeric features may take on large numbers of values (hundreds or thousands), so it is often desired to significantly reduce that number; for example, because large numbers of values may unnecessarily increase the subsequently generated model complexity. For humans, the task of analyzing features with so many values is very difficult, if not impossible. Grouping feature values into a smaller number of intervals eases the task. For example, age can take on integer values from 0 to 120 years, or even more values if age is measured in months or days. Thus, often age is discretized by grouping ages into intervals such as infants (0–1 year), children (1–12), teenagers (13–20), etc., or into decades (1–10; 11–20; and so on) depending on the goal of analysis.

Discretization can be performed in two ways: with or without taking into account class information, such as diagnosis. Discretization algorithms that do not use class information are called unsupervised; those that do are called supervised.

The goal of discretization is to reduce the number of values of a continuous feature by grouping them into a smaller number of intervals (bins). The main question is how to choose the number of intervals; a related question is how to set the interval widths. The number of intervals can be specified by a user, using some heuristic, or an algorithm can decide this number on its own by using class information.

Heuristics for discretization

One heuristic says that the number of intervals n_F for a feature F should not be smaller than the number of classes C . A second heuristic is given by

$$n_F = M / (3 \times C), \quad (20.1)$$

where M is the number of training instances. Remember that both of these heuristics require knowledge of class membership (diagnosis) for each data point.

Unsupervised discretization algorithms

The simplest unsupervised discretization algorithms are *equal width* and *equal frequency*. The equal width algorithm first finds the minimum and maximum values of the i th feature F_i and then divides the range between them into user-specified number of equal-width intervals. The equal frequency algorithm also finds the range, but then sorts all values in ascending order, and puts into each interval the same number of the sorted values. Thus, the interval widths vary when the equal frequency algorithm is used.

Supervised discretization using CAIM

A problem with discretization is that always some of the original information is lost. The supervised information-theoretic *class-attribute interdependency maximization* (CAIM) discretization algorithm uses the association between the feature being discretized and classes to accomplish discretization while minimizing loss of information about the interdependency defined by the CAIM criterion (Kurgan & Cios, 2004). Class information is used by the algorithm and as the result the users need not specify the number of intervals.

To explain how the CAIM algorithm works, we first define the quanta matrix (also called a contingency table) shown in Table 20.2. In the quanta matrix, rows represent classes (e.g., different diseases, or disease stages), columns are nonoverlapping intervals for values of the feature being discretized (e.g., continuously varying biomarker), each of the (s by n) cells contain frequencies, and the last row and column are the total frequencies for each row and column, respectively. The value of the CAIM criterion is calculated at each iteration step that divides the feature values into intervals, until an approximately optimal value of n is obtained.

Table 20.2 Quanta matrix

Class	Interval					Class Total
	$[d_0, d_1]$	...	$(d_{r-1}, d_r]$	...	$(d_{n-1}, d_n]$	
C_1	q_{11}	...	q_{1r}	...	q_{1n}	M_{1+}
$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
C_i	q_{i1}	...	q_{ir}	...	q_{in}	M_{i+}
$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
C_S	q_{S1}	...	q_{Sr}	...	q_{Sn}	M_{S+}
Interval Total	M_{+1}	...	M_{+r}	...	M_{+n}	M

Note: From Kurgan and Cios (2004).

Suppose F is a feature with many continuous values to be discretized. There is a discretization scheme D that places values for F into n mutually exclusive and exhaustive intervals bounded by the pairs of numbers

$$D : \{[d_0, d_1], (d_1, d_2], \dots, (d_{n-1}, d_n]\},$$

where d_0 and d_n are the minimal and maximal values of F , and the values are arranged in ascending order. Taken together, the values

$$\{d_0, d_1, d_2, \dots, d_{n-1}, d_n\}$$

constitute the boundary set (upper and lower values for each interval) for the discretization scheme D .

Each step of the CAIM algorithm divides one of the existing intervals into two new intervals, if the division improves the value of the CAIM criterion (see below). CAIM starts with the entire interval $[d_0, d_n]$ and maximizes interdependence between the continuous feature and its class labels to automatically generate a small number of discrete intervals for the feature F . The CAIM criterion measures dependency between the class variable C and the discretization variable D , for a given quanta matrix as

$$CAIM(C, D | F) = \frac{\sum_{r=1}^n \frac{\max_r^2}{M_{+r}}}{n} \quad (20.2)$$

where n is the number of intervals, r iterates through all intervals (i.e., $r = 1, 2, \dots, n$), $\max_r$ is the maximum value among all q_{ir} values (maximum in the r th column of the quanta matrix), $i = 1, 2, \dots, s$ (number of classes), and M_{+r} is the number of continuous values of attribute F that are within the interval $(d_{r-1}, d_r]$. Notice that the larger the number of values belonging to class C_i within a particular interval, the higher the interdependence between C_i and the interval. CAIM is maximized when all values for F within a particular interval belong to the same class, for all intervals. CAIM favors discretization schemes with small numbers of intervals.

CAIM consists of two steps: (a) initialization of the candidate interval boundaries and the corresponding initial discretization scheme, and (b) the consecutive additions of a new boundary that results in the locally highest value of the CAIM criterion; for details see Kurgan and Cios (2004). The CAIM algorithm's implementation in MATLAB can be downloaded from www.mathworks.com/matlabcentral/fileexchange/24344-caim-discretization-algorithm or from www.cioslab.vcu.edu/alg/main.html (and run in the WEKA environment following the provided instructions). Application and discussion of discretization of continuous features in clinical data sets can be found in (Maslove, Podchiyska, & Lowe (2013) and for the purpose of improving predictive modeling of disease outcomes in Hayward et al., (2010) are appearing in the health sciences literature.

Model building

After data are preprocessed using feature selection, discretization, and many other methods, model building may commence. Choice of a particular model building method depends heavily on characteristics of the available data and preferences of the users.

At the highest level of abstraction, supervised and unsupervised methods for model building are recognized. Unsupervised methods do not require knowledge of class labels for any data

instances. For example, we may have images of patient lungs but without any accompanying diagnosis. Supervised methods require class information for each data point; data that include class information are known as training data; for example, when images of lungs of each patient are accompanied by corresponding diagnosis (such as cancer type).

Unsupervised learning methods

Only few methods can be used for building models from unsupervised data (data without class information). The most important are clustering algorithms such as K-means, Kernel-based, and K-medoids (e.g., Kaufman & Rousseeuw, 1987; MacQueen, 1967). These methods find natural subgroups (clusters) in data. For example, lung image data could be grouped into clusters that may subsequently be found by experts to correspond to patients with small cell lung cancer, healthy patients, and patients with large cell lung cancer. Another key unsupervised method, called association rule mining, is used for finding associations between data points. Examples are algorithms such as Naïve algorithm (Fisher, 1936), Apriori (Agrawal & Srikant, 1994), and multilevel association rules (Piatetsky-Shapiro, 1991; Thakur et al., 2007). Such algorithms can find (associate) that patients with small cell lung cancer also have high blood pressure, for example.

Supervised learning

Some supervised machine learning algorithms generate models of data that are easy to understand. Of special interest are those algorithms that built a model of data as a set of rules in this format:

IF image has features “xyz” THEN small cell lung cancer.

Supervised learning algorithms are classified on the basis of type of training data available. Three scenarios are that (a) each data point belongs to one of the classes (single-instance learning); (b) groups of points (called bags) belong to one of the classes (multiple-instance learning; MIL); or (c) learning is based on data points belonging to one class only (one-class learning, or outlier detection). Each scenario will be described and issues related to their application will be considered. Rule-based approaches for scenarios (a) and (b) are presented next.

Scenarios for supervised learning

In the first, easiest scenario, each data point belongs to one of the classes. For example, we have training data about 150 patients. Each patient is diagnosed as having either small cell lung cancer (70 patients) or as being healthy (80 patients). If we generate model of the data in terms of IF . . . THEN . . . rules, the features used in the rules are the ones that best distinguish one group of patients from the other. The model can later be used to predict whether other patients have small cell lung cancer or are healthy based on their lung images. The training and test data for this scenario are shown in Figure 20.3. Artificial pulmonary function data are used to explain the idea. (Note that usually only some features extracted from images are used in practice, because using the entire images even as small as 1026×1026 pixels in size would mean using well over one million individual features/pixels.)

Multiple-instance learning (MIL) is more difficult. In MIL, data about different diagnoses based on data obtained during multiple visits over varying periods of time are available for

Data	ID	FVC	FEV	PEF	Class
Training	1	Low	Normal	Low	Yes
	2	Low	High	High	No
	3	High	Low	Normal	No
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	149	Low	Low	Normal	Yes
	150	Low	Low	Low	Yes
Test	New	Low	High	Low	?

Figure 20.3 Training and test example data sets. Training data are on the first 150 rows and test data (new patient) is shown at the bottom. The “?” indicates that the trained model needs to predict whether a new patient has small cell lung cancer or not. FVC = forced vital capacity; FEV = forced expiratory volume; PEF = peak expiratory flow; Class *Yes* = small cell lung cancer, *No* = no small cell lung cancer.

each patient (Dietterich, Lathrop, & Lozano-Perez, 1997; Nguyen, Nguyen, Hobson, Kurgan, & Cios, 2012). Each patient is treated as a “bag” with diagnoses made over time by physicians during patient visits. A bag is labeled “positive” if it has *at least one* positive instance (e.g., patient diagnosed with small cell lung cancer during one of the visits) and negative otherwise (during none of the visits a patient was diagnosed with small cell lung cancer). A bag (patient) thus may have a different number of diagnoses (instances). In a positive bag, the instances can be either true positive (patient diagnosed during one of the visits with small cell lung cancer) or false positive (the same patient possibly diagnosed during the visits with something other than small cell lung cancer diseases). In a negative bag, by contrast, all instances are true negatives (patient was never diagnosed with small cell lung cancer, but may have been diagnosed with other diseases).

Following on our example, the 150 patients in the training set are now seen as “bags” with data for the bags obtained during multiple visits. Each of the 70 patients (bags) sick with small cell lung cancer has been possibly diagnosed with many disease, but we know that at least one of them was small cell lung cancer. For the remaining 80 patients (bags) we know that none had small cell lung cancer (but possibly other diseases). This is obviously a more difficult model-building scenario. When the model is used for predicting a label for a new patient (bag), it can suggest that a patient has (or not) the small cell lung cancer. Table 20.3 illustrates definitions of a bag, positive bag, negative bag, true positive, and false positive used in the MIL scenario. Figure 20.4 shows exemplar training and test data for the MIL problem.

The third scenario, known as one-class learning (also called outlier detection) is more difficult yet because we have data representing only one class (say, of patients with one disease). As shown in Figure 20.5, we might have data about 70 patients, all diagnosed with small cell lung cancer, but no other data (like about healthy patients, or patients with other disease). Our task is to come up with a model to be used later for predicting that a new patient has (or not) small cell lung cancer.

Often data exist with a significant imbalance between the positive (with cancer) and negative (normal/healthy) cases and in some situations it is beneficial to redefine the binary classification problem as a one-class learning problem. For example, the binary classification

Table 20.3 Concepts used in MIL

Concept	Definition	# of instances in a bag	Features			Class/Bag Label
			FVC	FEV	PEF	
Bag	Collection of instances (diagnoses).	1	High	Normal	Normal	
		2	Low	High	High	
		3	Normal	High	High	
		4	Low	Low	Low	
Positive Bag	A bag with <i>at least one positive</i> instance. (Diagnoses 1 and 3 are True Positives, while 2 and 4 are False Positives.)	1	High	Normal	Normal	Yes (SCLC)
		2	Low	High	Low	
		3	Normal	High	High	
		4	Low	Low	Low	
Negative Bag	A bag with <i>no positive</i> instances. (All diagnoses are True Negatives.)	1	High	High	Normal	No (no SCLC)
		2	High	Normal	High	

Note: FVC = forced vital capacity; FEV = forced expiratory volume; PEF = peak expiratory flow; class/bag label = yes (small cell lung cancer) or no.

Data	Patient (Bag)	Instance	FVC	FEV	PEF	Class
Training	1	1	Low	Normal	Low	Yes
		2	High	Normal	Normal	
	2	1	High	Low	Normal	No
		2	Normal	Low	Normal	
		3	Normal	High	High	
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	149	1	Normal	Normal	Normal	No
	150	1	Low	High	Low	Yes
		2	Normal	Normal	Normal	
		3	Low	Medium	Low	
Test	New	1	High	Normal	Normal	?
		2	Low	High	Low	

Figure 20.4 Training and test example for multiple-instance learning. Training data appear toward the top, and test data are at the bottom. In the training data set, we do not know which particular instance of a patient (bag) with *Yes* in the Class column is the true positive (having SCLC); however, a *No* in the Class column means that all diagnoses are true negatives. In the test data set, we have new patient data and the MIL algorithm is used to predict whether the patient has SCLC or not, as indicated by the ?. FVC = forced vital capacity; FEV = forced expiratory volume; PEF = peak expiratory flow; Class *Yes* = small cell lung cancer, *No* = no small cell lung cancer.

Patient	FVC	FEV	PEF	Class SCLC
1	Low	Normal	Low	Yes
2	Low	Low	Low	Yes
.	.	.	.	Yes
.	.	.	.	Yes
.	.	.	.	Yes
69	Normal	Low	Normal	Yes
70	Normal	Low	Low	Yes
New	Low	High	Low	?

Figure 20.5 Data for a one-class learning problem. The training data set of 70 patients all have small cell lung carcinoma (SCLC = Yes). One-class learning algorithm is to determine whether test data instance (a new patient) also has SCLC (or not).

problem “What is the infection status of hospitalized patients (infected, not infected)?” could be translated to a one-class learning problem by instead asking “Is the patient not infected? The approach is to determine the boundary of the feature space enclosing the larger group known to be infection-free; if a new case falls outside the boundary, the case is regarded as an outlier that arose from some other class. Using this approach can improve classification performance. To illustrate, Cohen et al. (2004) reported a case where 11% of the cases were positive while the remaining 89% were negative. When this problem was solved using a binary (single-instance) classification approach, sensitivity was 50.6%. On the other hand, when solved as a one-class problem using only the majority class (89% of data) and discarding the minority class (11% of data) altogether, sensitivity increased to 92.6%. Thus, if one has hundreds of patient data with a certain disease, but only a small number of data from healthy people, it is beneficial to build a model using one-class learning approach and use the majority class (patients with disease) only. For more information about one-class learning algorithms, see Barnett and Lewis (1994) and Hempstalk, Frank, and Witten (2007).

Rule-based algorithms

Let us note that rules can be “written out” from a decision tree, but building a model in terms of rules is different. Rule algorithms generate rules that are independent, which means that a single rule can be understood without reference to other rules; this is not the case for rules written out from a decision tree (Quinlan, 1993). Many algorithms can generate models in terms of rules, with several freely available from WEKA software at www.cs.waikato.ac.nz/ml/weka (Witten, Frank, & Hall, 2011). We describe here two rule learning algorithms, one for the single-instance (Figure 20.3) and one for the MIL scenario (Table 20.3, Figure 20.4).

Below we first describe the DataSqueezer (DS) algorithm to represent algorithms for the first learning scenario (Kurgan, Cios, & Dick, 2006). Second, we describe the mi-DS algorithm (Nguyen et al., 2012) as an exemplar of algorithms of the MIL scenario.

DataSqueezer algorithm can be downloaded from <http://sci2s.ugr.es/keel/algorithms.php>. Both algorithms can be also downloaded from www.cioslab.vcu.edu/alg/main.html (and run in the WEKA environment following the provided instructions). The reason for choosing these

algorithms is that mi-DS is based on the DS algorithm so it is easier to explain and understand the concepts.

Single-instance learning scenario: DataSqueezer algorithm

The DataSqueezer algorithm for single-instance learning generates a model of data in terms of IF . . . THEN . . . rules that use original features to best distinguish, say, one group of patients from another. Figure 20.6 shows pseudocode for the two-step DS algorithm; first, data are reduced; then, rules are generated.

We illustrate how this algorithm works by using the toy example shown in Figure 20.7. Data are shown in Panel A. In Step 1, the algorithm performs data reduction to generalize information present in the data. Data reduction is performed for both positive (with disease/yes) and negative (normal/no) data examples, as shown in Panel B. It results in generation of the G_{POS} and G_{NEG} matrices/tables shown in Panel C, which include counts of the number of examples for each pattern in the original data (Panel A).

In Step 2, a rule is generated using a search procedure that starts with an empty rule and adds selectors (selectors are feature values) until the termination criterion is satisfied. The process is illustrated in Figure 20.8. The rule consists of selectors generated by checking the G_{POS} table against the G_{NEG} table. If the rule covers any data in G_{NEG} , a new selector is added to make it more specific to distinguish between positive and negative data rows. The maximum depth of the search is equal to the number of features. Next, the examples covered (recognized correctly) by the generated rule are removed from the POS table, and the process is repeated; for details see Kurgan, Cios, & Dick (2006).

Multiple-instance learning scenario: mi-DS algorithm

The mi-DS algorithm generates rules from training data consisting of bags of instances (patients diagnosed with many diseases) instead of single instances (patients diagnosed with just one

```

Input: Training data set
Step 1: Data Reduction
    Distribute training data into POS and NEG tables
    Reduce instances in both POS and NEG tables and generate  $G_{POS}$  and  $G_{NEG}$  tables
Step 2: Generate Rules
    Reset pattern  $P$  (all values of all features = *)
    RULES = empty, STOP = false
    While not STOP
        Choose the most suitable feature value to append into  $P$ 
        If ( $P$  does not match any row in  $G_{NEG}$ ) then
            Begin
                Add  $P$  to RULES
                Delete all row in  $G_{POS}$  that match  $P$ 
                Reset pattern  $P$ 
            End
        If  $G_{POS}$  = empty then STOP = true
    End While
Output: Set of RULES that cover all positive instances
  
```

Figure 20.6 Pseudocode of DataSqueezer algorithm.

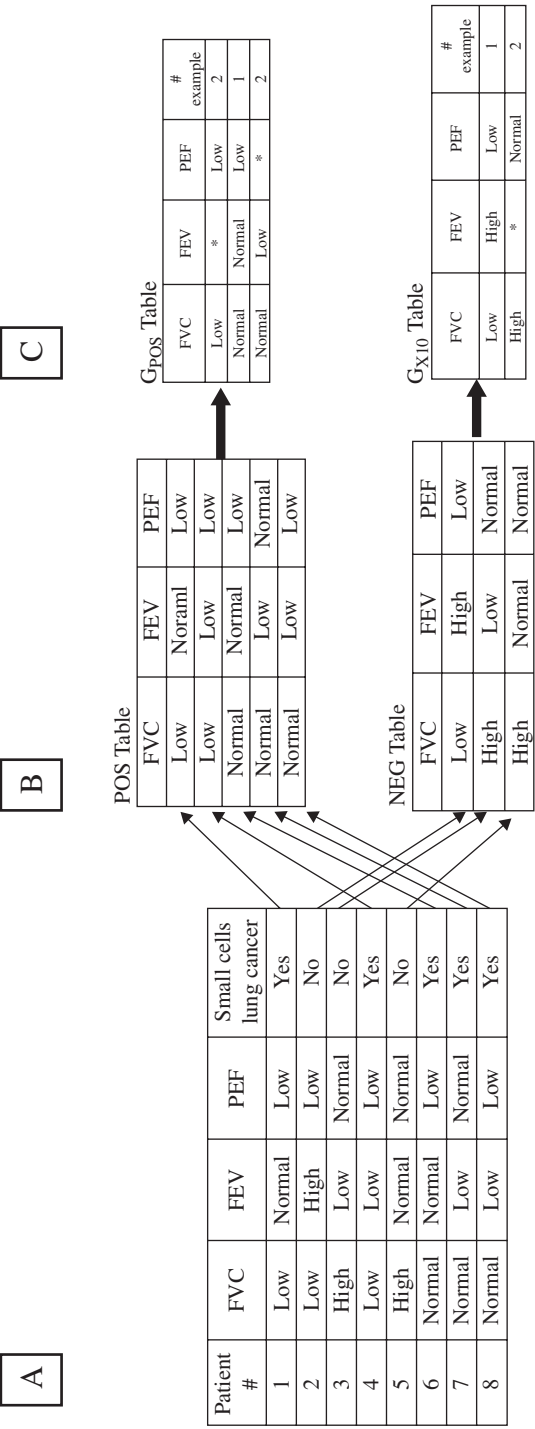


Figure 20.7 Toy example for the DataSqueezer algorithm step 1: Data Reduction. The complete data table shown in Panel A is first divided into separate tables for positive and negative instances of small cell lung cancer, shown in Panel B. then, as shown in Panel C, generalized feature patterns are identified and the number of examples that each row “covers” (matches) in the POS/NEG tables is counted; a “*” means “do not care” (what the particular value of a feature is)

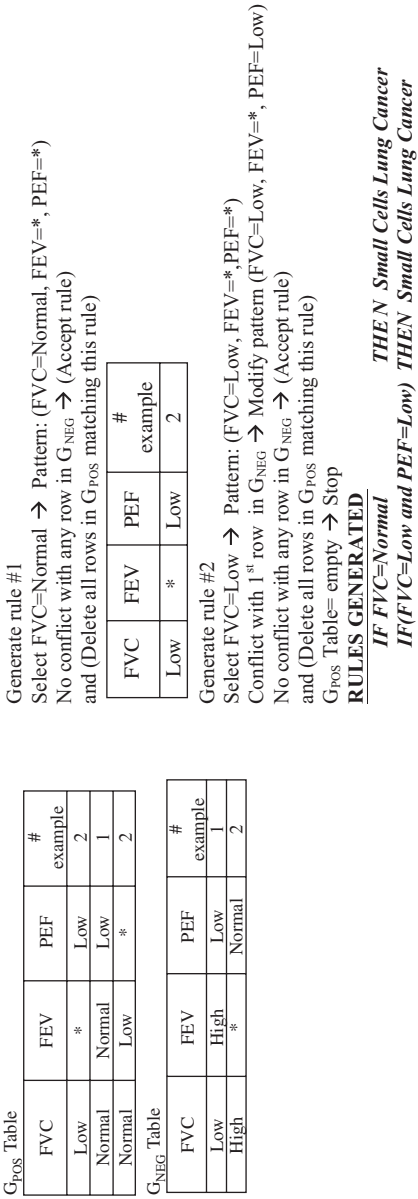


Figure 20.8 DataSqueezer algorithm, step 2: MakeRules. The G_{POS} and G_{NEG} tables are carried over from Figure 20.7, Panel C. The text shows how two rules that cover/identify cases of small cell lung cancer are generated

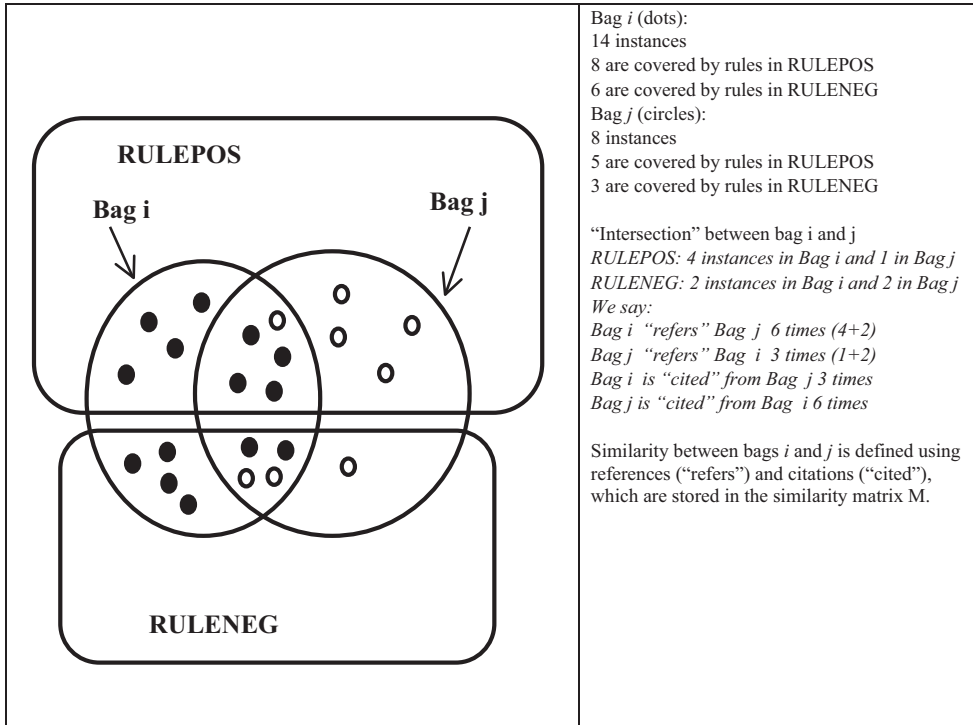


Figure 20.9 Illustration of the similarity between two bags.

disease). To clarify, in single-instance learning a patient is a “simple object” diagnosed with one disease only. In MIL, a patient is a “complex object” as it is diagnosed with many diseases, but only one of the diseases is of interest to us for the purpose of modeling; the specific disease we call “positive” (in our example, it is small cell lung cancer). Thus, there are positive bags (patients diagnosed with small cell lung cancer at least once) and negative bags (patients not diagnosed with small cell lung cancer during any of the visits). Defining the similarity measure between two bags is central to the operation of mi-DS algorithm. We explain how it is done through the means of an example shown in Figure 20.9.

Pseudocode of the mi-DS algorithm is shown in Figure 20.10. The label of a new test bag *B* (patient) is determined by calculating similarity between *B* and all the training bags; the concept is illustrated in Figure 20.11 but not its actual implementation. The similarity calculations are done using the similarity matrix, **M**, and the generated rules covering positive and negative training examples (stored in RULEPOS and RULENEG tables).

Figure 20.12 illustrates mi-DS using four bags (two positive and two negative) as input, and the generated four rules, shown in RULEPOS and RULENEG tables.

Figure 20.13 illustrates construction of the similarity matrix. In the left table, we show how many instances each rule covers in each bag for both RULEPOS and RULENEG. For example, in Bag₄ there are two instances covered by the first rule from the RULEPOS table and one instance covered by the second rule from the RULENEG table. Next, the similarity matrix is constructed as follows. The entry at location (row *i*, column *j*) is equal to S_{ij} (similarity between Bag_{*i*} and Bag_{*j*}). For example in the left table of Figure 20.12, we have two instances in Bag₂ and

Training phase Input: training bags (positive and negative bags) Distribute instances of positive bags into POS Table, instances of negative bags into NEG Table Delete instances in POS Table which also appear in the NEG Table Call DataSqueezer to create rules for all positive instances: RULEPOS Call DataSqueezer to create rules for all negative instances: RULENEG Construct similarity matrix M based on distance of the bags in the training set Output: RULEPOS for positive instances, RULENEG for negative instances, similarity matrix M
Testing phase Input: test bag B, RULEPOS, RULENEG, similarity matrix M Measure distances between test bag B with all training bags based on RULEPOS, RULENEG then update matrix M Build reference list (R_{list}) and citation list (C_{list}) of test bag B based on value R and C Predict class of test bag B based on all bags in R_{list} and C_{list} Output: predicted label of test bag B

Figure 20.10 Pseudocode for the multiple instance data squeezer (mi-DS) algorithm.

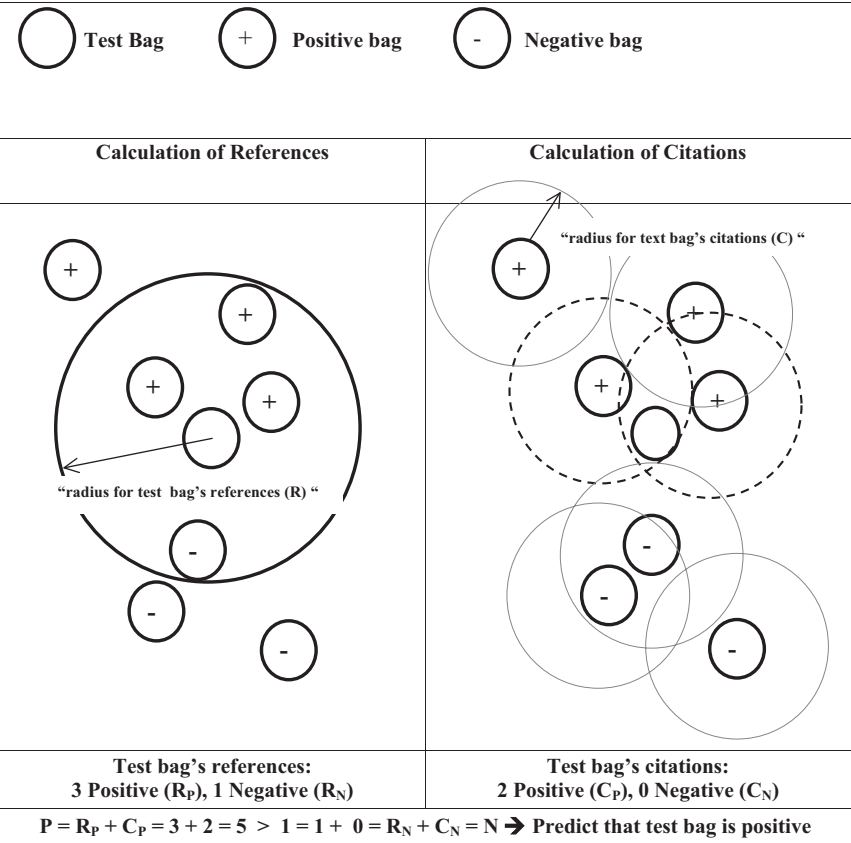


Figure 20.11 Determining the label for a new bag (patient) B.

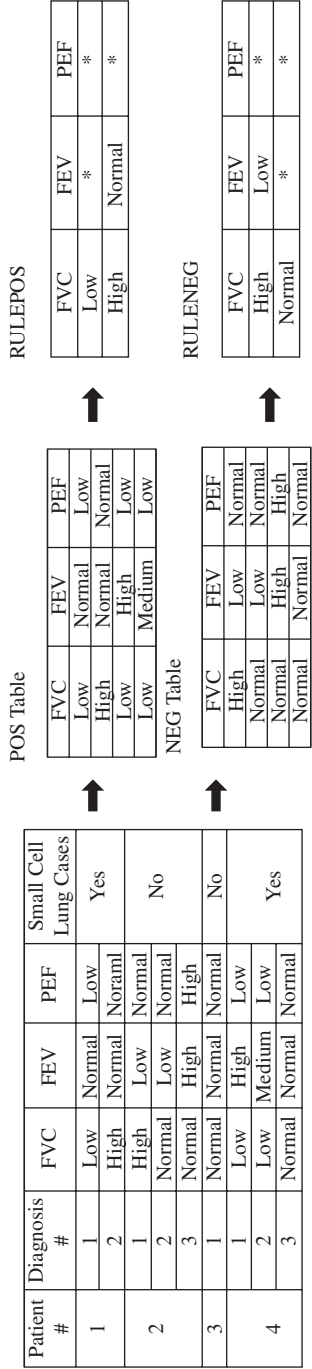


Figure 20.12 Rule generation using the mi-DS algorithm.

RULENEG	Low, *, *	1			2
	High, Normal, *	1			
	High, Low, *		1		
	Normal, *, *		2	1	1

→

Bag ₂	Bag ₃	Bag ₄
0	0	1
0	2	2
1	0	1
1	1	0

Figure 20.13 Construction of a similarity matrix.

New Patient (Bag)	Instance	FVC	FEV	PEF	Class SCLC
B	1	High	Normal	Normal	?
	2	Low	High	Low	
	3	Normal	High	High	
	4	Low	Low	Low	

Figure 20.14 New bag (patient) B data.

	Bag ₁	Bag ₂	Bag ₃	Bag ₄	B
Bag ₁	0	0	0	1	2
Bag ₂	0	0	2	2	2
Bag ₃	0	1	0	1	1
Bag ₄	2	1	1	0	3
B	3	1	1	3	0

$R_{list} = \{Bag_1(+), Bag_4(+)\}$
 $C_{list} = \{Bag_4(+), Bag_1(+), Bag_2(-)\}$
 $P (= R_p + C_p = 2 + 2 = 4) > N (N = 0 + 1 = R_N + C_N)$

 Diagnosis/prediction: Test bag B (patient B) has small cell lung cancer.

Figure 20.15 Prediction of the label for new patient/bag B.

one instance in Bag₄ covered by the rule “Normal,*,*” (see RULENEG table in Figure 20.12), so $S_{24} = 2$ and $S_{42} = 1$. The right table of Figure 20.13 shows the resulting similarity matrix.

To diagnose patient B (test bag B) shown in Figure 20.14, we first calculate similarities between bag B and all other bags. Having the similarities, we predict the class of test bag B using values $C = 2$ and $R = 2$, as illustrated in Figure 20.15. For details, refer to Nguyen, Nguyen, Hobson, Kurgan, and Cios (2013).

Validation

Once several models have been generated, a decision about which one is “best” is needed. After the best model is selected, it can be used for making predictions for new test data instances (patients) that were not used in generating the model. For practical examples see (Cudeck & Henly, 1991; Geisser, 1980; Linhart & Zucchini, 1986).

Identification of the best model is based on the predictive power (error of prediction) of the models. All models have some error associated with them; the errors are calculated as the difference between the observed (true) value and the model output (predicted) value. During model generation, we *fit the model* to the data. However, our main interest is in using the model for making *predictions* on new data. Thus, any generated model needs to be assessed for its *fit error* and its *prediction error*. In the machine learning literature, prediction error is often referred to as the generalization error. The generalization error terminology ties model validation to concepts of overfitting (model

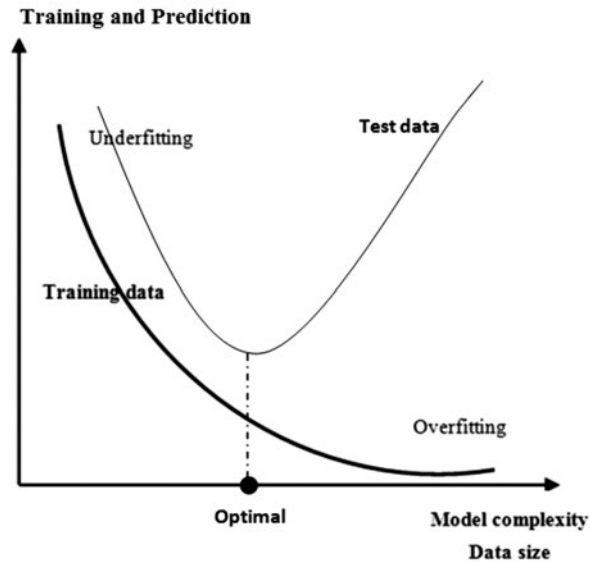


Figure 20.16 Choosing an optimally complex model with acceptable fit and prediction errors.

fits the training data too well, so that it will not generalize) and underfitting (model does not capture essential aspects of the data). Figure 20.16 illustrates these concepts. Below we describe key methods for selecting the best model when training data are available.

Training and testing data

The simplest validation method involves splitting the available (typically large) training data into the training part (used to fit the model) and the test part, used to evaluate prediction error. Typically about two-thirds of the data, chosen randomly, is used for training and the rest is used for testing.

Cross-validation

A popular and more accurate method is called *cross-validation*, usually used on medium sized and small data. In a k -fold cross-validation, we divide the entire data set, consisting of n points, into k equal-sized subsets, and use $k - 1$ parts for training and the remaining one part for calculating the prediction error. We repeat the procedure k times, and report the average from the k runs. A *10-fold cross-validation*, illustrated in Figure 20.17, is frequently used. When $k = 10$, the prediction error is $1/10$ times the sum of all ten calculated errors. For small data sets, the n -fold cross-validation is used, with $k = n$. The n -fold cross-validation approach is also known as the *leave-one-out* method.

Bootstrapping

Still another technique is called the bootstrap. Bootstrapping uses sampling with replacement to form the training data set. We sample a dataset of n instances n times *with replacement* to form a new dataset of n instances, and use this data as the training data set. (Note that due to the sampling with replacement, the bootstrapped sample may include duplicate instances.) Then we use

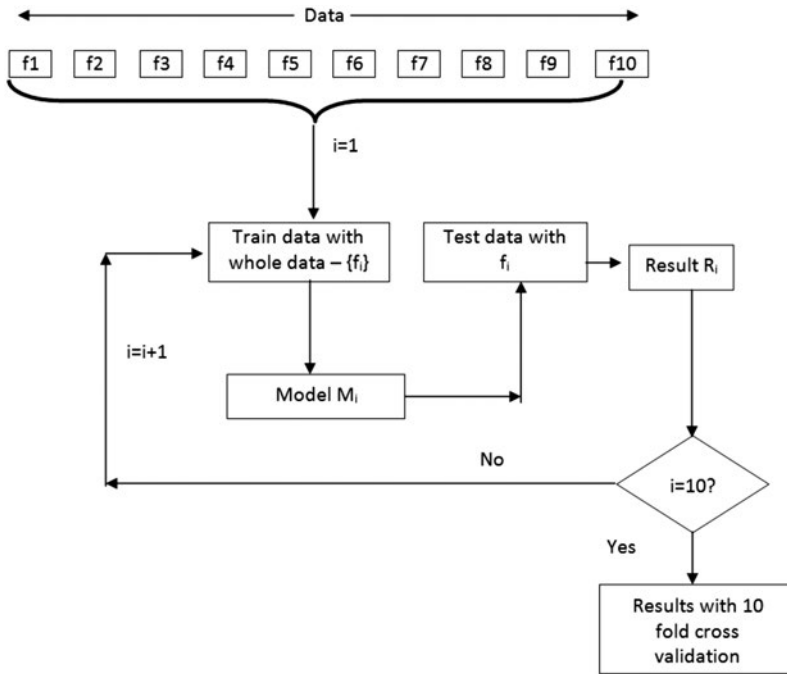


Figure 20.17 Illustration of the process for 10-fold cross-validation.

the instances from the original dataset that do not occur in the new training data set for testing. Thus, we fit the model to all bootstrapped samples, and calculate the prediction error using the original data. The problem with this approach is that prediction error is too optimistic (good), since the bootstrap samples are drawn with replacement. In other words, we calculate the prediction (generalization) error on overlapping data (many data items in the test set being the same as in the training set). There are several variations of bootstrap methods. A commonly used one is the .632 bootstrap where on average, 63.2% of the original data instances would end up in the training set and the remaining 36.8% would form the test set (Han & Kamber, 2006).

Classification errors

Classification errors are calculated based on information present in the training data, as we know class (diagnosis) for each data point (patient). In this case we can calculate errors committed by a generated model as differences between the actual outcomes and the predicted outcomes.

Binary classification

Since we know input and the corresponding output for each training data point, we can construct a *misclassification matrix* (also known as *confusion matrix*, or *contingency table*). The misclassification matrix for a binary classification problem is shown in Table 20.4. Note that a multiclass classification problem can always be decomposed into several binary classification problems, so the method is general. Notice that a model/classifier always makes some errors predicting either false negatives or false positives.

Table 20.4 Misclassification matrix

True Class ^a	Model/classifier-predicted Class		Total
	Positive	Negative	
Positive (P)	TP	FN (error)	P = TP + FN
Negative (N)	FP (error)	TN	N = FP + TN

Note: Cell entries are counts. TP = true positives; FN = false negatives.
P = all positives; FP = false positives. TN = true negatives; N = all negatives.

^a“True” class is obtained using some gold standard.

Table 20.5 Measures of classifier correctness

Index	Formula	Meaning
Sensitivity ^a	TP/(TP + FN)	Sensitivity is a classifier’s ability to correctly identify TP cases. It is the probability of a positive test given that the patient is truly sick. Sensitivity of 1 means that the model recognizes all truly sick people as sick.
Specificity	TN/(FP + TN)	Specificity is the ability of a classifier to correctly identify TN cases. It is the probability of a negative test given that the patient is truly healthy. Specificity of 1 means the model recognizes all healthy people (negative) as not sick.
False Positive Rate ^b	1 – specificity = FP/(FP + TN)	FPR defines how many incorrect positive cases occur among all truly negative cases.
Precision	TP/(TP + FP)	Proportion of TP to all predicted positives cases (thus includes both TP and FP).

Note: TP = true positive; FN = false negative; FP = false positive; TN = true negative; FPR = false positive rate.

^aSensitivity is also known as recall or true positive rate (TPR).

^bFalse positive rate (FPR) is also known as false alarm rate.

Suppose that positive (P) cases indicate sick people and negative (N) cases normal/healthy people. True positives (TP) are sick people correctly predicted as sick; false negatives (FN) are sick people incorrectly predicted as healthy (error); true negatives (TN) are healthy people correctly predicted as healthy; false positives (FP) are healthy people incorrectly predicted as sick (error). From the misclassification matrix we can calculate several types of errors, some of which are shown in Table 20.5.

Note that a FP diagnosis (with null hypothesis of being healthy) may cause a prescription of an unnecessary treatment. A FN diagnosis, on the other hand, can give a patient the illusion of good health (and the patient will not get the necessary treatment). Let us use the numbers from our example, so that P = 70 (sick) and N = 80 (healthy), and assume that TP = 60 and TN = 60 (thus FP = 20 and FN = 10); then

$$\text{Sensitivity} = \text{TPR} = \text{TP}/(\text{TP} + \text{FN}) = \text{TP}/P = 60/70 = 0.857$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) = \text{TN}/N = 60/80 = 0.75$$

$$\text{False Positive Rate (FPR)} = 1 - \text{Specificity} = 1 - 0.75 = 0.25$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) = 60/(60 + 20) = 0.75$$

Receiver operating characteristic curves

Receiver operating characteristic (ROC) curves are another popular technique for model evaluation. The ROC space is defined by FPR shown on the x -axis, and true positive rate (TPR; sensitivity) shown on the y -axis. It depicts a trade-off between TP (benefits) and FP (costs). One point in the ROC space corresponds to just one of the misclassification matrices associated with a specific model/classifier. An example ROC plot that compares goodness of two classifiers A and B is shown in Figure 20.18.

As shown in Figure 20.18, a few misclassification matrices, each of which give rise to only one point on the curve, are needed to approximate the curve for each classifier.

In general the points (and the corresponding misclassification matrices) for the ROC curve are easily generated by nondiscrete classifiers, such as neural networks with a sigmoidal function at the output node that produce a continuous output; using different thresholds on this output results in obtaining different TP and FP numbers and thus misclassification matrices. Figure 20.19 illustrates the idea. For the threshold of 3.5, there are corresponding TP, FP, TN, and FN numbers, which corresponds to just one ROC point. Using a threshold of 5, gives rise to the second ROC point, and so on.

Having generated two ROC curves for classifiers A and B, we can decide which of the two classifiers is better. Visually, the curve more towards the upper left indicates a better classifier, as the point $(0, 1)$ represents an ideal classifier with $\text{TPR} = 1$ and $\text{FPR} = 0$. Since visual analysis can be misleading, the area under the curve (AUC) is used to quantify classification quality using a single number summary; the larger the AUC, the better the classifier/model. Based on the ROC curves in Figure 20.18, it is easy to notice that classifier B is better than classifier A.

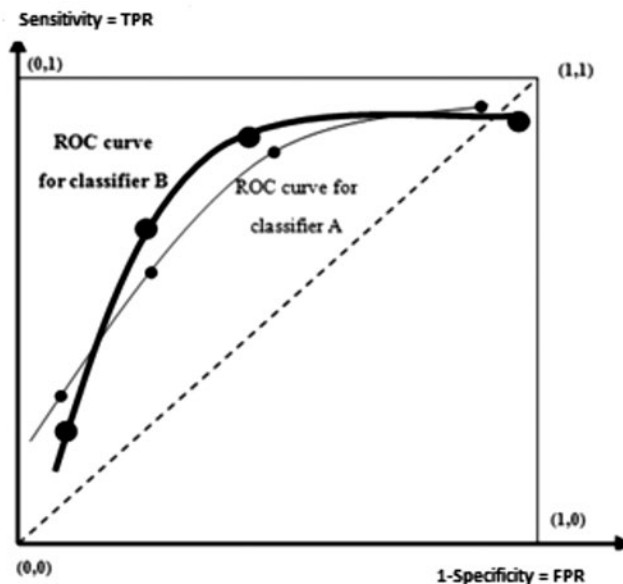


Figure 20.18 Receiver operating curve (ROC) comparing performance of classifiers A and B.

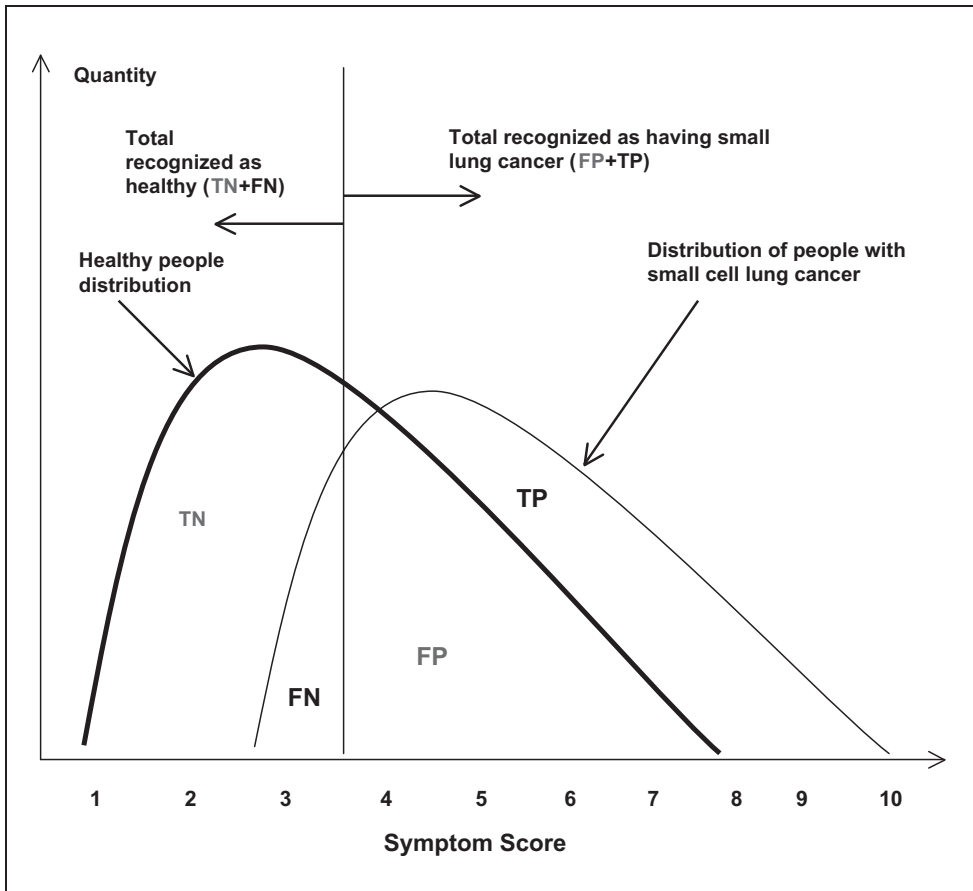


Figure 20.19 Threshold impact on misclassification.

Visualization

In many areas of medicine, biology, and engineering, there is often lack of background knowledge (diagnosis) about data, like having only undiagnosed images of patient lungs. In those situations, clustering is the most important technique for revealing structure in the data. The goal is to find clusters, if they exist. With further analysis, experts could label clusters as sick patients or healthy people, for example.

Unfortunately clustering is difficult, as it depends on many user-specified parameters, especially when data are highly dimensional. The key parameter is determining a similarity measure to be used for grouping the data points. Most algorithms also require that a user specifies a priori a number (usually only an educated guess) of clusters to be found in the data. Note that there can be between 2 and $n - 1$ (where n is the total number of points) clusters in data.

An alternative approach is to visualize the higher-dimensional data in a low-dimensional space, such as 3-D. Regions in this space with high density of data points would reveal the data's internal structure (number of clusters). To be useful, an algorithm mapping data from the highly

dimensional space into the low-dimensional space has to be faithful/accurate in the sense of preserving the most important characteristics of the original data. Usually, fidelity of such a mapping is achieved by using a similarity measure that preserves the property that if data points are close (or distant) in the original space they are also close (or distant) in the 3-D space. Below we describe three such visualization methods.

Sammon's projection

Sammon's projection (Sammon, 1969) is an algorithm that maps high-dimensional data into a space of lower dimensionality. This nonlinear projection attempts to preserve topological relations between patterns in the original and the reduced lower-dimensional spaces by preserving the inter-pattern distances. The following error function, referred to as Sammon's stress, E , is minimized

$$E = \frac{1}{\sum_{i < j} d_{ij}^*} \sum_{i < j} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*}, \quad (20.3)$$

where d_{ij}^* is distance between i th and j th points in original space while distance between their projections is denoted by d_{ij} . The minimization can be performed either by gradient descent or by involving some iterative methods. The number of iterations needs to be experimentally determined and convergence is not guaranteed, which are weak points of the algorithm.

Kohonen's self-organizing map (SOM) algorithm

The concept of a self-organizing feature map (SOM) was introduced by Kohonen (1982). SOM can be regarded as a two-dimensional neural network, composed of a grid/number of artificial neurons, that visualizes highly dimensional data in a 2-D structure/map. Again, to make visualization faithful to original patterns in the data, the requirement is that the 2-D representation preserves topological properties of the data. There are many variations of the SOM, but the generic architecture and learning algorithm are basically the same. The input is a set of multidimensional vectors and each neuron in the grid has associated with it a weight vector of the same dimensionality as the input vector. The SOM learning consists of initializing the neuron weights, presenting the input data, finding the "winning" neuron for each input, and adjusting weights of the "winning" neuron and its neighbors. A pseudocode of the SOM is shown in Figure 20.20.

Input: Data set of M instances, N features
 Initialize weights
 For 1 to total number of training steps
 Select a sample from the input data set
 Find the "winning" neuron for the sample input
 Adjust the weights of the winning neurons and its neighbors
 End
Output: Show 2D-graph of M points in $\mathbb{R}^2$ space

Figure 20.20 Pseudocode of the self-organizing map (SOM) algorithm.

Molecular dynamics method

Visualization using the molecular dynamics method is based on a technique simulating dynamics of atoms and molecules. It calculates time-dependent behavior of a particle system and to study the structure, dynamics, and thermodynamics of particle ensembles. The main steps of the method are: (a) defining a system consisting of M particles, and an initial set of parameters; (b) solving Newton's equations of motion for each particle, until the thermodynamic state of the system reaches equilibrium; and (c) recording positions and moments of particles in the equilibrium – the final state of the system.

The goal is to conserve all distances between points in original $\mathbb{R}^N$ space in the projected 3-D or 2-D Euclidean space, with minimum error. The method was adopted for data projection and visualization in the following way (Dzwinel & Błasiak, 1999; Nguyen, Dzwinel, & Cios, 2011). Every particle in the $\mathbb{R}^3$ space corresponds to a data point in $\mathbb{R}^N$. The particles i and j interact via two-body “potential” V_{ij} , which is a function of distance D_{ij} between the points i and j in the original space. See Figure 20.21. Simulations are run starting from some initial configuration of particles until equilibrium of the system is achieved. At each time step, the transient forces acting on each particle are computed and the particles are shifted to new positions. The particles stop moving when the potential energy reaches minimum (indication of equilibrium state). Figure 20.22 shows the pseudocode of the algorithm.

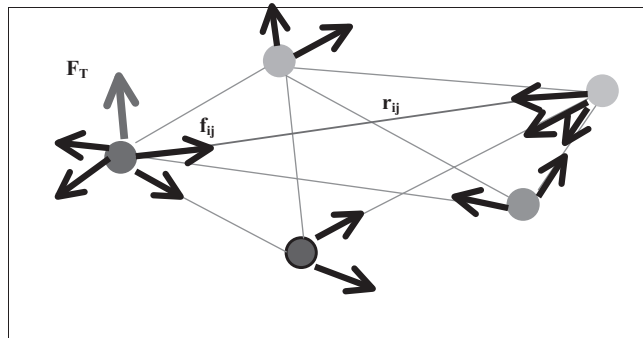


Figure 20.21 Illustration of interacting particles/points in 2-D space.

Input: Data set with M instances, N features

Initialize system with M points in $\mathbb{R}^3$ space corresponding with M points in $\mathbb{R}^N$ (M instances, N features) while the system is in motion

for $i = 1$ to $M - 1$ do // calculate Forces

for $j = i + 1$ to M do

Calculate force between particle i and j

Update total force F_i on particle i

Update total force F_j on particle j

end

end

for $i = 1$ to M do Update velocity and position of particle i based on F_i

end

Output: Show 3D-graph of M points in $\mathbb{R}^3$ space

Figure 20.22 Pseudocode for the molecular dynamics algorithm.

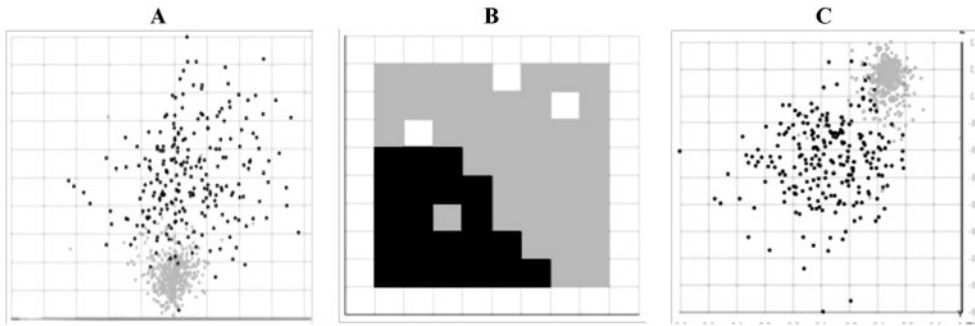


Figure 20.23 Visualization of breast cancer data.

Example for visualization approaches

We use the Breast Cancer Wisconsin data to illustrate how Sammon's projection, the SOM, and molecular dynamics methods can be used to visualize highly dimensional data in a low-dimensional space. The data set consists of 699 instances, described by 10 features, which belong to two classes (malignant breast cancer or benign breast lesion). Figure 20.23 shows the visualizations. Notice that in this example, the reduction is from 10-D space to 2-D space, and that all three algorithms show quite good separation of the data in the 2-D space. The separation into two easily visible groups in both the Sammon's (Panel A) and molecular dynamics (Panel C) projections is quite good. However, the separation is not so apparent in the SOM result (Panel B). In fact, SOM's projections are generally difficult to interpret and can be done only when we know class memberships for at least a few data points. The data set is available from the UCI Machine Learning Repository at: <http://archive.ics.uci.edu/ml/datasets.html>. Excellent implementations of the SOM algorithm are available in MATLAB or can be downloaded from Helsinki Institute of Technology at www.cis.hut.fi/projects/somtoolbox/. All three programs, and the data used in Figure 20.23, can be also downloaded from www.cioslab.vcu.edu/alg/main.html.

Software

Although we provide many references, very good information relating to many aspects of data mining can be found at www.kdnuggets.com run by Gregory Piatetsky-Shapiro. Readers should be aware that although there are many commercial and open-source data mining tools available they do not produce good results after simply clicking "run" as we need to understand their applicability and limitations. Validation is critical, as results may be completely wrong even when seeming to be very good.

References

- Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*. Santiago, Chile. Retrieved from <http://www.rsrikant.com/papers/vldb94.pdf>
- Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. Chichester, West Sussex, England: Wiley.
- Berger, A. M., & Berger, C. R. (2004). Data mining as a tool for research and knowledge development in nursing. *CIN: Computers, Informatics, Nursing*, 22, 123–131.

- Cios, K. J., & Kurgan, L. (2005). Trends in data mining and knowledge discovery. In N. R. Pal & L. Jain (Eds.), *Advanced techniques in data mining and knowledge discovery*, (pp. 1–26). London: Springer.
- Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, 26, 1–24.
- Cios, K. J., Pedrycz, W., & Swiniarski, R. (1998). *Data mining methods for knowledge discovery*. Boston, MA: Kluwer Academic.
- Cios, K. J., Pedrycz, W., Swiniarski, R., & Kurgan L. (2007). *Data mining: A knowledge discovery approach*. Heidelberg, Germany: Springer.
- Cohen, G., Hilario, M., Sax, H., Hugonnet, S., Pellegrini, C., & Geissbuhler, A. (2004). An application of one-class support vector machines to nosocomial infection detection. In M. Fieschi (Ed.), *MedINFO 2004. (Studies in health technology and informatics)*, 107(Pt. 1), 716–720. Amsterdam, the Netherlands: IOS Press.
- Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and “problem” of sample size: A clarification. *Psychological Bulletin*, 109, 512–519.
- Dietterich, T. G., Lathrop, R. H., & Lozano-Perez, T. (1997). Solving the multi-instance problem with axis-parallel rectangles. *Artificial Intelligence Journal*, 89, 31–71.
- Dzwiniel, W., & Błasiak, J. (1999). Method of particles in visual clustering of multi-dimensional and large data sets. *Future Generation Computer Systems*, 15, 365–379.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge discovery and data mining: Towards a unifying framework. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (pp. 82–88). Portland, OR.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, Part II, 179–188.
- Geisser, S. (1980). A predictivistic primer. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics* (pp. 363–381). Amsterdam, the Netherlands: North Holland.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- Hall, M. A. (1999). Correlation-based feature subset selection for machine learning. Unpublished doctoral dissertation. The University of Waikato, Hamilton, New Zealand. Retrieved from www.cs.waikato.ac.nz/~mhall/thesis.pdf
- Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann.
- Hand, D. T. (1998). Data mining: Statistics and more? *American Statistician*, 52, 112–118.
- Hand, D. T., & Bolton, R. J. (2004). Pattern discovery and detection: A unified statistical methodology. *Journal of Applied Statistics*, 31, 885–924.
- Hayward, J., Alvarez, S. A., Ruiz, C., Sullivan, M., Tseng, J., & Whalen, G. (2010). Machine learning of clinical performance in a pancreatic cancer database. *Artificial Intelligence in Medicine*, 49, 187–195.
- Hempstalk, K., Frank, E., & Witten, I. H. (2007). One-class classification by combining density and class probability estimation. Department of Computer Science, University of Waikato, Hamilton, NZ. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.8501&rep=rep1&type=pdf>
- Huber, D., Schumacher, L., & Delaney, C. (1997). Nursing Management Minimum Data Set (NMMDS). *Journal of Nursing Administration*, 27, 42–48.
- Hylock, R., Street, W. N., Lu, D.-F., & Currim, F. (2008). NursingCareWare: Warehousing for nursing care research and knowledge discovery. In J. Li, D. Aleman, & R. Sikora (Eds.), *Proceedings of the 3rd INFORMS Workshop on Data Mining and Health Informatics* (pp. 1–6). Retrieved from http://dollar.biz.uiowa.edu/~street/research/informs08_hylock.pdf
- Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. In Y. Dodge (Ed.), *Statistical data analysis based on the L_1 norm and related methods* (pp. 405–416). Amsterdam, the Netherlands: North-Holland.
- Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., & Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7, 81–99.
- Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Ninth International Workshop on Machine Learning* (pp. 249–256). San Francisco, CA: Morgan Kaufmann.
- Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2), 273–324.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59–69.

- Kudo, M., & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33, 25–41.
- Kurgan, L. A., & Cios, K. J. (2004). CAIM discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 16(2), 145–153.
- Kurgan, L. A., Cios, K. J., & Dick, S. (2006). Highly scalable and robust rule learner: Performance evaluation and comparison. *IEEE Systems, Man, and Cybernetics – Part B: Cybernetics*, 36(1), 32–53.
- Linhart, H., & Zucchini, W. (1986). *Model selection*. New York, NY: Wiley.
- MacKenzie, S. L., Wyatt, M. C., Schuff, R., Tenenbaum, J. D., & Anderson, N. (2012). Practices and perspective on building integrated data repositories: Results from a 2010 CTSA survey. *Journal of the Medical Informatics Association*, 19, e119–e124.
- MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). Berkeley: University of California Press.
- Maslove, D. M., Podchiyska, T., & Lowe, H. J. (2013). Discretization of continuous features in clinical datasets. *Journal of the American Medical Informatics Association*, 20, 544–553.
- National Science Foundation. (n.d.). Core techniques and technologies for advancing big data science and engineering (BIGDATA). Retrieved from www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767&org=CISE
- Nguyen, D. T., Dzwinel, W., & Cios, K. J. (2011, November). Visualization of highly-dimensional data in 3D space. In *Intelligent Systems Design and Applications Conference* (pp. 225–230). Córdoba, Spain.
- Nguyen, D. T., Nguyen, C. D., Hobson, R., Kurgan, L. A., & Cios, K. J. (2012). mi-DS: Multiple instance learning algorithm. *SMC-Part B*, 43(1), 143–150.
- Nguyen, D. T., Nguyen, C. D., Hobson, R., Kurgan, L. A., & Cios, K. J. (2013). mi-DS: Multiple instance learning algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 43, 143–154.
- Pal, N. R., & Jain, L. C. (2005). *Advanced techniques in knowledge discovery and data mining*. London, UK: Springer-Verlag.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11), 559–572.
- Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro, W. J. Frawley & J. William (Eds.), *Knowledge discovery in databases* (pp. 229–248). Cambridge, MA: AAAI/MIT Press.
- Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann.
- Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18, 401–409.
- Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4), 13–19.
- Thakur, R. S., Jain, R. C., & Pardasani, K. R. (2007). Fast algorithm for mining multilevel association rule mining. *Journal of Computer Science*, 1, 76–81.
- Westra, B. L., Savik, K., Oancea, C., Choromanski, L., Holmes, J. H., & Bliss, D. (2011). Predicting improvement in urinary and bowel incontinence for home health patients using electronic health record data. *Journal of Wound, Ostomy and Continence Nursing*, 38, 77–87.
- Westra, B. L., Subramanian, A., Hart, C. M., Matney, S. A., Wilson, P. S., Huff, S. M., & Delaney, C. W. (2010). Achieving “meaningful use” of electronic health records through the integration of the Nursing Management Minimum Data Set. *Journal of Nursing Administration*, 40, 336–343.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining. Practical machine learning tools and techniques* (3rd ed.). San Francisco, CA: Morgan Kaufmann.
- Witten, I. H., Frank, E., & Hall M. A. (2011). Data mining practical machine learning tools and techniques [Software]. Available from www.cs.waikato.ac.nz/ml/weka/index.html
- Woodside, A. G., Nielsen, R., Walters, F., & Muller, G. D. (1988). Preference segmentation of health care services: The old-fashioned, value conscious, affluent, and professional want-it-all. *Journal of Health Care Marketing*, 8, 14–24.
- Wu, X., Zhu, X., Wu, G., & Ding, W. (2013). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, Issue: 99, doi:10.1109/TKDE.2013.109

21

GENOMIC, TRANSCRIPTOMIC, EPIGENOMIC, AND PROTEOMIC APPROACHES

Yvette P. Conley

Variation in health and disease in human beings is related in part to molecular processes at the cellular level due to variation in DNA sequence (genomics), variation in gene expression (transcriptomics), variation in non-sequence related changes in DNA (epigenomics), and variation in protein production (proteomics) within the cell. It is therefore often justified to incorporate investigation of these omic approaches into a program of nursing research.

The purpose of this chapter is to summarize key approaches and design considerations for genomic, transcriptomic, epigenomic, and proteomic studies. Association studies using DNA sequence variability, gene expression studies, epigenomic studies, and proteomic studies will be addressed. The objectives are to prepare readers to understand and evaluate omic studies and to motivate consideration of incorporating omic approaches into appropriate content areas of nursing research. Table 21.1 lists internet resources related to the issues that are addressed.

Scientific concepts

Genome and related concepts

The genome refers to all of the inherited material found in the cell of an organism. For a human, this material is packaged into 23 pairs of chromosomes in the nucleus as well as the genetic material found in the mitochondria. The terms genetic and genomic are often used interchangeably; however, they are not identical. Genetic refers to focus on a particular gene while genomic refers to focus on the entire genome (all genes and their interactions). The difference is scale, where “omic” refers to a larger scaled study (e.g., transcriptomic, proteomic).

To place the approaches discussed in this chapter the reader should keep in mind the central dogma of molecular biology: a unit of DNA (i.e., gene) houses the sequence code for an amino acid chain (i.e., a protein); the DNA acts as a template for an intermediate template called messenger RNA (mRNA) through a process called transcription; and the mRNA acts as a template for the amino acid chain through a process called translation. Therefore DNA acts as a template for mRNA that is a template for an amino acid chain that becomes a protein.

A polymorphism is a particular spot in the DNA that is variable in the population and occurs at a frequency in the population of at least 1%. Alleles are alternate forms of a particular spot in the DNA; for example, if a polymorphism has two alleles in the population that means that

Table 21.1 Internet resources for genomic, transcriptomic, epigenomic, and proteomic approaches

<i>Topic/Study Type</i>	<i>Content</i>	<i>Site</i>
General Background	Talking glossary of genetic terms	www.genome.gov/glossary
	Online Mendelian Inheritance in Man (OMIM)	www.ncbi.nlm.nih.gov/omim
	Quanto – sample size calculator	http://hydra.usc.edu/gxe
	CaTS – power calculator	www.sph.umich.edu/csg/abecasis/cats/
	Genetic Power Calculator	http://pngu.mgh.harvard.edu/~purcell/gpc/
	Patient Reported Outcomes Measurement Information System (PROMIS)	www.nihpromis.org
	Consensus Measures for Phenotypes and Exposures (PhenX)	www.phenx.org
Gene Expression	Gene Expression Omnibus	www.ncbi.nlm.nih.gov/geo/
	Gene Ontology Database	www.geneontology.org/
	European Bioinformatics Institute	www.ebi.ac.uk/
	Serial analysis of gene expression (SAGE) database	www.sagenet.org
Genetic/Genomic Association	Catalog of published genome-wide association studies	www.genome.gov/gwastudies
	SNP database	www.ncbi.nlm.nih.gov/projects/SNP/
	International HapMap Project	http://hapmap.ncbi.nlm.nih.gov/
	Genes, Environment and Health Initiative	www.genome.gov/19518663
Epigenomics	Chromatin structure and function	www.chromatin.us
	Database for DNA Methylation and Environmental Epigenetic Effects	www.methdbe.de
	Database of noncoding RNAs	www.noncode.org
	Epigenomic datasets	www.ncbi.nlm.nih.gov/epigenomics
	Epigenomic Fact Sheet	www.genome.gov/27532724
	Histone Database	www.research.nhgri.nih.gov/histones
	Human Epigenome Project (HEP)	www.epigenome.org
	International Human Epigenome Consortium	http://ihec-epigenome.org
	MicroRNA data resource (miRBase)	www.mirbase.org
Proteomic Studies	Protein database	www.ncbi.nlm.nih.gov/protein
	Entrez Protein Clusters	www.ncbi.nlm.nih.gov/proteinclusters
	PRIDE: Proteomics IDentifications Database	www.ebi.ac.uk/pride

there are two alternative forms of that particular spot in the DNA. Alleles are called major or wild type for the most frequent allele and minor or variant for the least frequent allele. Some polymorphisms are functional and can change an amino acid sequence in protein in a way that impacts the function of that protein; it could impact a regulatory element for a gene so that transcription of that gene and hence amount of protein produced is altered, for example. Other polymorphisms are not functional but nevertheless serve a purpose in association studies because they can characterize a piece of DNA that is used in an analyses to see if that piece of DNA is associated with a phenotype of interest.

Within the context of this chapter, assay refers to chemical analysis that allows for qualitative or quantitative measure of a template (assay). Assays that have been ordered and fixed to a solid support that allow for simultaneous measure of a template are arrays. The collection of data from a large number of assays or arrays in a short period of time using technology and automation is referred to as high throughput data. The National Human Genome Research Institute's Talking Glossary contains definitions for many terms used in omics research (see Table 21.1).

Phenotype

Phenotype refers to an observable or measurable trait of an individual. Phenotypes are impacted by genetic makeup, environmental exposures, or interactions between both. Examples of phenotypes arising in nursing research are development of symptoms or complications within the context of chronic conditions, development and healing of chronic wounds, stress response in caregivers, and complex behavioral manifestations.

Phenotype data collection is an important consideration that should not be underestimated. The best designed omics study is only as good as the phenotype data that are utilized. Additionally, cases (those who exhibit the phenotype) and controls (those who do not exhibit the phenotype) both need to be consistently phenotyped as part of a well-designed study. Failure to replicate findings from omics studies is often due to variation in phenotype characterization. Consistency of phenotyping is the goal of the Patient Reported Outcomes Measurement Information System (PROMIS) and the Common Measures for Phenotypes and Genotypes (PhenX) initiatives. PROMIS and PhenX are excellent resources to use when determining how one might collect phenotype data. Additionally, reporting the details of how subjects were phenotyped is important when disseminating results from omics approaches so that replication of findings can be achieved.

Procedural issues

Recruitment

Recruitment scheme is important to consider whether initiating a study or using existing data and samples. Phenotype of interest can impact study design and recruitment options. A familial approach that recruits related individuals and investigates how variation in omics data segregates with a phenotype of interest is powerful. However, many phenotypes of interest, such as development of symptoms or complications within the context of a disorder, susceptibility to a complex condition where the number of affected family members may be small, response to an intervention, and recovery following an injury or debilitating disorder would not lend themselves well to a familial approach. Study design for these phenotypes will likely center around recruitment of unrelated individuals who share risk of developing the phenotype of interest with case control study designs most commonly utilized.

Biospecimens

Studies that focus on DNA sequence variability rely upon DNA as the template to perform data collection, which is an advantage compared to some of the other approaches addressed in this chapter. Advantages include that DNA is very stable and DNA for this approach can be extracted from any tissue at any time. These advantages mean that stored samples are very likely to be amenable to this approach.

The initial template used to collect data for DNA polymorphism-based and mRNA-based expression studies (DNA versus mRNA) is an important difference to consider. When conducting a DNA polymorphism-based study, one is not concerned with tissue of origin and timing of the sample collection and usually one is not worried about stability of DNA from stored samples. This is not the case with an mRNA-based study. Different tissues express different genes at different levels and gene expression can be dynamic and change over time; therefore, much consideration needs to be given to tissue of origin and timing of sample collection for mRNA. The tissue of origin and timing of sample collection for an expression study should represent the phenotype being studied as much as possible. As an example, if one were interested in what genes are up regulated (increased gene expression) or down regulated (decreased gene expression) for a cancer that presents with a solid tumor, mRNA extracted from the tumor could be compared to the normal tissue surrounding it. It is also possible to compare gene expression between individuals with a phenotype compared to those who do not have the phenotype of interest, but again what tissue is compared between these groups needs to be given consideration.

Blood is one of the most readily available tissues for most studies. In gene expression studies, an argument can often be made for use of blood as an acceptable tissue. This is particularly true if there is good rationale that there will be a systemic impact by the phenotype that may be represented in the blood; if proteomic studies of blood have found variability in the protein profile for that phenotype and now focus is on the mRNA level; or if one potential deliverable from the research is translation of findings for clinical use.

RNA instability needs to be considered when conducting an mRNA-based gene expression study. Whereas DNA is very stable, mRNA is not and it will degrade if not stabilized immediately after sample collection. Adding to the degradation concern is the issue that not all mRNA templates degrade at the same rate so data collected from even partially degraded samples will not be accurate. Fortunately, advances in technology have made the task of RNA stabilization less burdensome than in the past, when immediate processing for RNA extraction or immediate snap freezing were always required. These options are still available; however, products are now available to immediately stabilize RNA in solid tissue and body fluids, which greatly facilitates sample collection from a variety of settings and batch processing. Great consideration needs to be given to whether stored samples would be appropriate for a gene expression study due to the issue of RNA stability. If stored samples that have not been stabilized are being considered for use, storage history of the sample needs to be taken into account (for example has it gone through any freeze thaw cycles, was it immediately snap frozen). It is also recommended that a subset of samples be selected for RNA extraction to assess the quality and yield of RNA from the stored samples before moving forward with using stored samples for a gene expression study.

Epigenomic approaches share design considerations of polymorphism-based and gene expression-based studies. Epigenomic approaches are similar to polymorphism-based studies because the template of interest is DNA; therefore, stability is not an issue. Similarly, epigenomic chemical modifications are also usually stable. Like gene expression-based approaches, attention needs to be given to the origin of tissue because chemical modifications to DNA vary from tissue to tissue (explaining much of the difference in tissue to tissue gene regulation). Rationale should

be proved for use of blood as a tissue under study. In addition, timing of sample collection matters in epigenomic studies. For example, in a polymorphism-based study of symptoms developed after onset of a chronic condition, DNA could be collected at any time before, during, or after onset of symptoms. This is not the case for an epigenomic study, where it is necessary to both study DNA at the time(s) justified for the phenotype of interest and to ensure that the time point(s) of sample collection are consistent for each subject. Stored samples will often provide the quality and yield of DNA needed for an epigenomic study; however, tissue of origin and timing of sample collection of the stored samples need much consideration. This is particularly the case if samples were collected from different tissues; for example, some from blood and some from saliva, or without regard for timing of sample collection. These issues would not have been a concern in a DNA polymorphism-based study.

Association studies using DNA sequence variability

Genetic and genomic association studies rely upon variability in the DNA sequence as an independent variable(s). Genetic and genomic association studies are distinguished by the extent of genome covered by the study. Genetic studies usually refer to a focus on one gene, whereas genomic studies focus on more than one gene and often involve looking at gene-gene and gene-environment interactions. The variability in the DNA sequence data that can be collected and used for this type of study includes single nucleotide polymorphisms (SNPs), insertions/deletions of nucleotides, and repetitive DNA polymorphisms such as variable number of repeat polymorphisms (VNTRs), short tandem repeats (STRs), and copy number variants (CNVs).

Association studies are used when variability in the DNA sequence of a gene(s) may be related to a phenotype of interest. A candidate gene or candidate pathway association study requires a priori selection of a candidate gene or pathway of genes. A genome-wide association study (GWAS) investigates DNA sequence variability across the entire genome and therefore does not require a priori selections.

Candidate gene association study

A candidate gene approach is taken when a specific gene or set of genes is hypothesized to be involved with the phenotype of interest. Three key activities are: (a) identification of candidate gene(s), (b) generating genotype data, and (c) conducting the association analysis.

Identification of candidate gene(s)

Justification of the a priori selected genes is critical. The argument for focusing on the selected genes should be based on biological plausibility for the association within the context of the phenotype. What are the functions of the gene and its products that make it a good candidate for the phenotype under study? Potential candidate genes should be considered in context with other potential candidate genes. If there is good biological plausibility for a gene, think about what might be important upstream to the gene-created effects, as well as downstream effects. For example, a candidate gene may be selected because its product (mRNA or protein) makes sense for the phenotype of interest. When an enzyme metabolizes that gene product into its active form – the gene for that enzyme could be an important candidate to consider. When the gene product from a candidate gene acts through a cell surface receptor – the gene for that cell surface receptor could be an important candidate to consider. When the gene product from a candidate gene acts as a transcription factor for other genes – those other genes could be important candidates to consider.

When considering what might be a good candidate gene, it often helps to think about what is currently known about the phenotype, including what therapies are successfully utilized. For example, medications that impact the phenotype suggest biological pathways the medication works through, whether it interacts with a receptor, and whether the medication itself is a gene product.

Additionally, evidence from previous studies can provide rationale for candidate genes. If a protein has been implicated in the phenotype, then the gene for that protein may be a good candidate. Also, results from genome-wide evaluations (linkage, association, methylation, gene expression) can implicate a gene or region of the genome and provide evidence to support selection of particular candidate genes.

Generating genotype data

Polymorphism data are used to determine if there is an association between the gene(s) and phenotype under study. Selecting polymorphisms usually takes several things into consideration. The literature should be consulted to determine if there have been any functional polymorphisms identified for the gene. It often makes sense to include these in an evaluation, although focusing solely on functional polymorphisms for a gene will usually not fully evaluate all of the variability in a gene. To do this, tagging SNPs (tSNP) for the gene are selected. The idea behind tagging SNPs is it is not necessary to genotype every SNP known for a gene because SNPs are often linked to one another in the gene; these SNPs are said to be in linkage disequilibrium (LD). In this situation, knowledge of the genotype of an individual for one SNP allows generation of the genotypes for the linked SNPs around it. Selecting tSNPs is facilitated by using databases that contain LD data for SNPs across the genome (for example, HapMap; see Table 21.1). When selecting tSNPs, going beyond the coding region of the gene and out into the 5' and 3' untranslated regions is recommended in order to capture regulatory elements for the gene.

Association analysis in candidate gene studies

Genotype data are then used to conduct association analyses. Typically studies will conduct allele and genotype association-based analyses and if more than one polymorphism is investigated it is recommended that haplotype analyses be conducted. Power to detect an association if it exists will depend on many of the same things that need to be taken into account with any association study. Other power and sample size related issues to consider with a candidate gene association study include the extent to which you think genetic variation plays in the phenotype of interest (for some phenotypes this can be estimated using heritability measures) and allele frequencies for the polymorphisms, which will determine how often you expect to see a given genotype. Some software packages have the ability to perform power and sample size calculations for genetic association studies as well as some stand alone, free-to-download calculators such as Quanto and the Genetic Power Calculator (see Table 21.1).

Another issue that is important to consider when conducting any genetic or genomic association study is population stratification. Population stratification takes into account the ancestry of the individuals in a study. Allele frequencies for polymorphisms often vary significantly among populations from different ancestries. If population stratification is not taken into consideration, a significant difference in allele or genotype frequency that results in a significant association could be simply due to ancestry-based frequency differences and not due to an association with the phenotype. This results in a false positive finding. Use of ancestral informative marker data in the analyses may reduce the influence of population stratification.

Genome-wide association studies (GWAS)

A GWAS is undertaken when, instead of focusing on a gene or set of genes that are a priori selected, data are collected that allow for assessment of the entire genome (genome-wide). In GWASs, genotype data are collected for polymorphisms spaced throughout the genome, and the genotype data are analyzed for associations with the phenotype of interest, usually by comparing cases to controls. Significant associations are flagged for follow up. The location of each polymorphism included in a GWAS is known; therefore, when a significant association is found it is possible to determine where in the genome the association exists. Basically, GWAS let the data identify associations. This is an important advantage of taking a GWAS approach because many phenotypes of interest are not well understood biologically. Using a GWAS approach eliminates the need for a priori selection of biologically plausible genes and allows opportunity to discover novel biological insights into the phenotype of interest. Key activities related to a GWAS are how to collect the data, sample size and cost considerations, the potential utility of existing GWAS data resources, and conducting the association analyses.

Data collection, sample size, and cost

Why wouldn't everyone consider the GWAS approach, given how little is known about most phenotypes? Possible downsides to taking a GWAS approach are sample size requirements and costs. The number of polymorphisms required to genotype if covering the entire genome is substantially larger than what is used for candidate approaches.

Most investigators taking a GWAS approach will use readily available pre-designed arrays containing highly informative polymorphisms that are spaced appropriately to capture variability across the genome. These arrays usually contain 500,000 to higher than two million polymorphisms. Although all of the polymorphisms may not be completely independent from one another, one can imagine that to correct for multiple testing of this large number of polymorphisms, samples sizes will need to be sufficiently large to be powered with correction for multiple testing incorporated. Some software packages have the ability to perform power and sample size calculations for a GWAS as well as some stand alone, free-to-download calculators such as CaTS (see table of internet resources). Many statistical geneticists advocate for sample sizes of at least 1,000–2,000 for cases and the same number for controls. These numbers may be attainable for some phenotypes, but for many phenotypes the traditional GWAS approach and traditional analyses may not be an option. Additionally, given the larger sample size needed, the use of genome wide data collection arrays, and the more extensive data management and analyses required, a GWAS approach is much more costly than candidate approaches, although the cost of conducting a GWAS continues to decline as technology and software to conduct these studies continue to improve.

Alternative sources of GWAS data

One important additional research opportunity to mention within the context of GWAS is the availability to use databases that house GWAS data already collected by other studies to conduct analyses for a phenotype of interest. Online information is available for GWAS studies that have made their data available (both genotype as well as phenotype data) to the scientific community for further data analyses; see internet resources in Table 21.1. The databases can be searched by phenotype of interest (for example, specific cancers, cardiovascular disease or risk factors for cardiovascular disease, laboratory biomarkers, symptom or complication development, quality of life)

and study of interest (for example, the Framingham Heart Study, the Nurses' Health Study, the Jackson Heart Study) to determine what data from what study you may be able to request access to, that could be analyzed to support or add direction to your program of research.

GWAS association analysis

Many of the same considerations mentioned previously for candidate gene analysis also hold for GWAS association analysis. GWAS association analysis does differ, however, not just in the amount of data analyzed but also in the sophistication of the analyses. For example, such analyses can take into account the biological connections between genes. Specialized software has been developed to facilitate these more sophisticated analyses as well as assist with data cleaning and data quality checking. Although candidate gene association analyses resemble other more “traditional” and familiar association analyses, it is recommended that association analysis of GWAS data be conducted by individuals well versed in the management of GWAS data and use of the specialized software to fully extract the biological information embedded within the data.

Gene expression studies

A gene expression study is undertaken when variability in how actively a gene is transcribed is hypothesized to be involved with the phenotype of interest. These types of studies do not focus on DNA, but instead focus on the mRNA produced from a gene during transcription. An mRNA-based gene expression approach should be used when one hypothesizes that changes in mRNA levels including level of alternately spliced mRNAs (think variety of mRNAs from a gene) and level of mRNA for micro RNAs may play a role in the phenotype of interest.

Although all genes are represented in the DNA of all cells, not all genes are active (expressed) to the same extent in all cells. What genes are active depends on the needs of the cell and can change in response to exposures and environment. Gene expression differences can also occur within the context of a phenotype of interest and changes in gene expression may precede detectable clinical phenotypes.

Like DNA-based association studies, mRNA-based studies can focus on a priori selected candidates or use a genome-wide approach. Many of the design considerations for a candidate gene or a genome-wide mRNA-based expression study are the same as those for DNA-based association studies. Candidate approaches will require justification of selected candidates as well as selection and/or design of an assay that in this case investigates the piece of mRNA of interest. Genome-wide approaches will likely use a readily available pre-designed array that will measure the level of expression of mRNA representing genes across the entire genome. Because the number of different polymorphisms in the genome far outnumber the number of mRNAs that can be generated from a genome, genome-wide expression studies generate less data compared to a GWAS (on average, 47,000 versus over two million data points, respectively per subject); however, the technology to collect these data and the density of genome-wide data for both association and gene expression data are constantly changing.

Epigenomics

The epigenome is made up of the chemical compounds that modify (“mark”) the genome in a way that affects gene expression. Thus, an epigenomic study is undertaken to evaluate variability in chemical modification or packaging of the DNA that ultimately results in changes to gene expression that may be involved with the phenotype of interest.

Epigenomic mechanisms

The chemical and DNA packaging modifications that occur to the DNA are important mechanisms behind the regulation of some genes, can be dynamic and change in response to exposures from the endogenous or exogenous environment, and can be sustained through cell division and passed on to future cells. The fact that chemical and DNA packaging modifications to the DNA are a mechanism to adjust gene regulation in response to environmental exposures makes it a very appealing addition to gene-environment studies. Studies characterizing the appropriate environmental exposures involved with a phenotype, especially if environmental exposure data and biospecimens are available over time, are particularly strong studies to introduce an epigenomic component.

Epigenomics is a rapidly changing field of inquiry. Currently, epigenomic mechanisms may involve histone modification, chromatin remodeling, noncoding micro RNA (miRNA), and DNA methylation. Histones are proteins essential to packaging DNA; these proteins undergo posttranslational modifications that impact its packaging ability and influence whether a gene will be active or repressed. Chromatin refers to the combination of the DNA plus all of its packaging-related proteins. How loosely or tightly packaged the chromatin is in a region of the genome plays a role in how up regulated or down regulated gene expression will be, respectively, for genes in that region. Noncoding miRNA are short single stranded pieces of RNA coded for by genes that are epigenetically regulated themselves; miRNA do not act as a template for translation into a protein but instead interact with mRNAs, rendering them unable to be translated and essentially silencing the gene. DNA methylation involves addition of methyl groups to cytosines within cytosine-guanine (CG) rich regions of DNA (CpG islands); hypermethylation results in that gene not being transcribed. These epigenomic mechanisms are important to gene regulation in response to needs of the cell or tissue, often in response to the environment.

Epigenomic studies

Epigenetic studies focus on a priori selected candidate genes, whereas epigenomic studies focus on part or all of the genome. Many design considerations for candidate gene or genome-wide epigenomic studies are the same as those mentioned previously for DNA- and RNA-based association studies. Candidate approaches require justification of selected candidates and selection and/or design of an assay that evaluates the region of the genome of interest. Genome-wide approaches, particularly for genome-wide evaluation of DNA methylation, likely use readily available pre-designed arrays. Arrays currently available that measure the extent of DNA methylation across the entire genome will usually generate around 27,000 to 450,000 data points; the technology to collect these data and the density of genome-wide CpG islands, shelves, and shores are constantly being updated.

Proteomic studies

A proteomic study is undertaken when varying levels of protein (end result of an expressed gene) are thought to play a role in a phenotype of interest. Protein-based studies can focus on an a priori selected protein, a related set of candidate proteins, or through advances in data collection technologies, evaluate all proteins in a sample. Candidate protein(s) selected for investigation require scientific justification and selection of an assay that investigates the protein(s) of interest. Moving from low throughput antibody-based protein quantification methods to evaluation of the proteome is made possible by improvements in technology to differentiate and quantify

proteins en masse. Readily available pre-designed arrays are available to assess the proteome or select pathway related proteins in ways that are similar to a whole genome polymorphism, gene expression, or epigenomic evaluation. Additionally, interrogation by high pressure liquid chromatography and mass spectrum technologies allow for proteome assessment that is not reliant on existing information that is required when using antibodies or an array-based method. With this approach, novel proteins or protein isoforms may be identified.

Proteomic approaches share many design considerations with gene expression-based studies. Stability of the template, tissue of origin, and timing of sample collection all need to be considered. Although the template of interest is protein and not mRNA, stability is still an issue. Like mRNA, different proteins will have varying stability and will degrade over time so attention needs to be given to stabilizing protein immediately upon collection and proper storage for downstream uses. In addition to the more traditional stabilization methods, such as snap freezing and use of additives like glycerol, use of advances in protein stabilization that allow for stabilization of the protein from aqueous as well as solid tissue can increase utility of the samples. Some of these advances are the same as those used for RNA stabilization and are commercially available. Great consideration needs to be given to whether stored samples would be appropriate for a protein-based study. The storage history of the sample needs to be taken into account and it is recommended that a subset of samples be selected for protein extraction to assess the quality, yield, and utility of the protein from the stored samples before moving forward with a proteomic study.

Different tissues express different genes at different levels; therefore, what proteins are made will differ between different tissues at different times. Much consideration needs to be given to tissue of origin of the protein. The tissue of origin for a protein-based study should represent the phenotype being studied as much as possible, and selection of that tissue should be justified, including use of blood. For example, comparing protein levels between individuals with a phenotype to those who do not have the phenotype of interest requires that the protein be extracted from the same phenotypically relevant tissue in cases and controls, collected at a consistent time point within and between these groups. Like gene expression and epigenomic-based studies, protein levels can be dynamic and change even within an individual over time and in response to environmental exposures; therefore, justification of the timing of repeated within subject assessments is a critical design component for all three of these approaches. Stored samples may provide the quality and yield of protein needed; however, tissue of origin and timing of sample collection of the stored samples will need to be given much consideration also. This is particularly the case if samples were collected from different tissues or without regard for timing of sample collection.

Conclusions

Genomics, transcriptomics, epigenomics, and proteomic approaches are increasingly being incorporated into research trajectories across disciplines. It is important to point out that while genomics, transcriptomics, epigenomics, and proteomic approaches are useful in helping to identify individuals at risk for a phenotype, this is not the only utility of these approaches. They are also excellent tools to help us better understand the pathophysiology of a phenotype, which is essential for the development of evidence-based interventions and understanding variability in response to interventions. It is also essential for progress towards personalized health care with the goal of better outcomes for patients. Nurse scientists are encouraged to add these biological approaches to their research trajectory with the goal of increasing understanding of phenotypes under investigation. Investigators would be remiss if they did not at least contemplate how addition of such approaches would add to the state of the science in their area of research.

This chapter was intended to introduce some of the more frequently utilized omic approaches and to encourage consideration of incorporating these approaches into their research program. This chapter was not meant to cover methods for data collection and analyses with details, mostly because of the rapidity in which the technologies for data collection and data analysis are changing. Instead, the chapter presents logistical issues to consider and provides information of a more conceptual nature. Resources such as seminal literature and internet resources are provided to assist the interested reader with a more in depth understanding of the concepts presented.

References and resources

Table 21.1 contains a list of internet resources that is organized by topic and content and provides addresses for the sites. The internet resources are from reliable sources and most importantly are kept up to date; therefore, utilization of these internet resources for the purpose of enhancing the experience of this chapter in as up to date a fashion as possible is encouraged.

References are also provided. The references contain seminal reads that provide a foundation for the approaches presented as well as literature that supports the logistics presented throughout the chapter. The central dogma of molecular biology was first described by Crick (1970). References for general information include Baumgartel et al. (2011) and Conley et al. (2013). Genetic and genomic association studies are described in Hong and Park (2012); Jorgensen et al. (2009); Kim et al. (2010); Le Marchand and Wilkens (2008); Liang, Thomas, and Conti (2012); Manichaikul et al. (2012); Nsengimana and Bishop (2012); Ritchie (2012); Sale, Mychaleckyj, and Chen (2009); and Sampson, Jacobs, Yeager, Chanock, and Chatterjee (2011). Gene expression studies are discussed in Bustin (2000); Coppola (2011); Livak and Schmittgen (2011); and Lovén et al. (2012). References for epigenomic studies include Cortessis et al. (2012); Heyn and Esteller (2012); Petronis (2010); and Rakyan, Down, Balding, and Beck (2011). Proteomic studies are discussed in Mischak et al. (2007); Nyangoma, Collins, Altman, Johnson, and Billingham (2012); Rodriquez-Suárez and Whetton (2013); and Tuñón et al. (2010).

References

- Baumgartel, K., Zelazny, J., Timcheck, T., Snyder, C., Bell, M., & Conley, Y.P. (2011). Molecular genomic research designs. *Annual Review of Nursing Research*, 29, 1–26.
- Bustin, S.A. (2000). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology*, 25, 169–193.
- Conley, Y.P., Biesecker, L.G., Gonsalves, S., Merkle, C.J., Kirk, M., & Aouizerat, B.E. (2013). Current and emerging technology approaches in genomics. *Journal of Nursing Scholarship*, 45, 5–14.
- Coppola, G. (2011). Designing, performing, and interpreting a microarray-based gene expression study. *Methods in Molecular Biology*, 793, 417–439.
- Cortessis, V.K., Thomas, D.C., Levine, A.J., Breton, C.V., Mack, T.M., Siegmund, K.D., & Laird, P.W. (2012). Environmental epigenetics: Prospects for studying epigenetic mediation of exposure-response relationships. *Human Genetics*, 131, 1565–1589.
- Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227, 561–563.
- Heyn, H., & Esteller, M. (2012). DNA methylation profiling in the clinic: Applications and challenges. *Nature Reviews Genetics*, 13, 679–692.
- Hong, E.P., & Park, J.W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics and Informatics*, 10, 117–122.
- Jorgensen, T.J., Ruczinski, I., Kessing, B., Smith, M.W., Shugart, Y.Y., & Alberg, A.J. (2009). Hypothesis-driven candidate gene association studies: Practical design and analytical considerations. *American Journal of Epidemiology*, 170, 986–993.
- Kim, S.Y., Li, Y., Guo, Y., Li, R., Holmqvist, J., Hansen, T., Pedersen, O., . . . Nielsen, R. (2010). Design of association studies with pooled or un-pooled next-generation sequencing data. *Genetic Epidemiology*, 34, 479–491.

- Le Marchand, L., & Wilkens, L. R. (2008). Design considerations for genomic association studies: Importance of gene-environment interactions. *Cancer Epidemiology*, 17, 263–267.
- Liang, W. E., Thomas, D. C., & Conti, D. V. (2012). Analysis and optimal design for association studies using next-generation sequencing with case-control pools. *Genetic Epidemiology*, 36, 870–881.
- Livak, K. J., & Schmittgen, T. D. (2011). Analysis of relative gene expression data using real-time quantitative PCR and the 2⁻(Delta Delta C(T)) method. *Methods*, 25, 402–408.
- Lovén, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., . . . Young, Y. A. (2012). Revisiting global gene expression analysis. *Cell*, 151, 476–482.
- Manichaikul, A., Chen, W. M., Williams, K., Wong, Q., Sale, M. M., Pankow, J. S., . . . Mychaleckyj, J. C. (2012). Analysis of family- and population-based samples in cohort genome-wide association studies. *Human Genetics*, 131, 275–287.
- Mischak, H., Apweiler, R., Banks, R. E., Conaway, M., Coon, J., . . . Yamamoto, T. (2007). Clinical proteomics: A need to define the field and to begin to set adequate standards. *Proteomics Clinical Applications*, 1, 148–156.
- Nsengimana, J., & Bishop, D. T. (2012). Design considerations for genetic linkage and association studies. *Methods in Molecular Biology*, 850, 237–262.
- Nyangoma, S. O., Collins, S. I., Altman, D. G., Johnson, P., & Billingham, L. J. (2012). Sample size calculations for designing clinical proteomic profiling studies using mass spectrometry. *Statistical Applications in Genetics and Molecular Biology*, 11(3) (Article 2).
- Petronis, A. (2010). Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature*, 465(7299), 721–727.
- Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12, 529–541.
- Ritchie, M. D. (2012). The success of pharmacogenomics in moving genetic association studies from bench to bedside: Study design and implementation of precision medicine in the post-GWAS era. *Human Genetics*, 131, 1615–1626.
- Rodriguez-Suárez, E., & Whetton, A. D. (2013). The application of quantification techniques in proteomics for biomedical research. *Mass Spectrometry Reviews*, 32, 1–26.
- Sale, M. M., Mychaleckyj, J. C., & Chen, W. M. (2009). Planning and executing a genome wide association study (GWAS). *Methods in Molecular Biology*, 590, 403–418.
- Sampson, J., Jacobs, K., Yeager, M., Chanock, S., & Chatterjee, N. (2011). Efficient study design for next generation sequencing. *Genetic Epidemiology*, 35, 269–277.
- Tuñón, J., Martín-Ventura, J. L., Blanco-Colio, L. M., Lorenzo, O., López, J. A., & Egido, J. (2010). Proteomic strategies in the search of new biomarkers in atherothrombosis. *Journal of the American College of Cardiology*, 55, 2009–2016.

A SURVEY OF THE SOURCES OF NOISE IN fMRI¹

*Douglas N. Greve, Gregory G. Brown, Bryon A. Mueller,
Gary H. Glover, and Thomas T. Liu*

Functional Magnetic Resonance Imaging (fMRI) is a tool for studying brain function, i.e., which neural systems are the brain substrate for a particular behavior. fMRI is noninvasive, meaning that neither surgery nor ionizing radiation are used to generate the images. It can also localize activity to a few millimeters. For these reasons, fMRI as a research area has exploded over the last 10 years. In task-based analysis,² the brain is treated as a black box to which a stimulus is applied and from which a response is measured. The fMRI measurement itself is related to the amount of deoxygenated hemoglobin in the blood and so it is called the Blood Oxygen Level Dependent (BOLD) signal. While dependent upon many other factors, the BOLD signal gives an indication of how much blood is flowing to a particular location in the brain at a particular time. Behavioral and thought processes cause neurons in a small area to become active, which increases blood flow to that area to supply the metabolic demand of the neural activity. The more intense the neural activation, the larger the increase in blood flow will be (though the relationship is not necessarily linear or straightforward). The analysis strategy is to correlate the BOLD waveform with the known time course of stimulation. Brain areas engaged in the task will have a BOLD signal that is correlated with the task; brain areas that are not engaged in the task will be uncorrelated. This allows the characterization of the neural substrate of behavior in terms of the locations in the brain that are engaged, the amount of brain tissue recruited, and the strength or amplitude of the response. This may be clinically relevant for an individual person for diagnosis, for treatment planning, for monitoring treatment response, or for assessing disease course. We are also interested in how populations vary, e.g., how do the location, intensity, and size of neural activation change between those diagnosed with schizophrenia versus those who are deemed clinically healthy? This leads to potentially four levels of analysis: (a) time series analysis of an individual at a particular visit date, (b) longitudinal analysis (i.e., from visit to visit), (c) between subjects within a defined population, and (d) between defined populations. When studying large samples, it is often necessary to scan them at different study locations, which can add a fifth analysis level (site). Each of these levels introduces a set of noise that will increase variability and/or introduce bias to the final result. This chapter describes how fMRI is analyzed at each of these levels, the sources of noise at each level, and ways to control the noise. We start with a case study to familiarize the reader with the terminology of fMRI as well as provide motivation for conducting a study. Next, we describe the analysis at the first and higher

levels based on the general linear model (GLM). The biophysics of fMRI are then summarized to introduce the reader to the physics of the measurement and a foundation for understanding the sources of the noise. Finally, the sources of noise at the various levels are surveyed along with methods to reduce their impact.

A case study

To provide a concrete example of how fMRI analysis is performed, we present a design from one of the studies performed by the Function Biomedical Informatics Research Network (fBIRN). The study used a working memory paradigm to explore how emotionally disturbing images interfere with working memory and whether this interference differs between healthy individuals and persons with schizophrenia. Subjects were asked to perform a working memory task that was divided into three phases. During the Encode phase, the subject was shown a 16-second series of pictures each with a single drawing and asked to save these in memory for recall 16–32 seconds later. These were followed by a 16-second Distractor phase in which images were either emotionally neutral (e.g., a chair) or emotionally disturbing (e.g., a car accident). In the Probe phase, the subject was shown a 16-second series of images with two drawings and asked to respond with a button press as to which one was in the original series. This entire sequence (called a “block”) lasted 48 seconds and gave five “conditions”: (a) Encode, (b) Emotional Distractor (ED), (c) Probe following Emotional Distractor (PED), (d) Neutral Distractor (ND), and (e) Probe following Neutral Distractor (PND). We are interested in brain areas that respond to each of these conditions individually as well as differentially. During a “run,” four such blocks were presented with 16 seconds between blocks during which scrambled, uninterpretable images were presented. The total run time was of 284 seconds. See Figure 22.1 for a diagram of the timing. This process was repeated seven more times for a total of eight runs during the visit. Multiple repetitions were used for averaging. All of this took place inside an MRI scanner. The subject lays on a bed inside of the bore of the scanner with his/her head inside of a coil apparatus used to detect the MRI signal. Images were presented to the subject either through goggles or projected onto a screen that the subject can see through a mirror mounted on the coil. Finger press responses were collected from a button box inside the scanner. The stimulus was controlled and button presses recorded by a single computer synchronized with the scanner.

During the time that images were being presented and button presses recorded, the MRI scanner was collecting images of the entire brain at a temporal sample rate of one every two seconds (called the “TR”). Each sample is referred to as a “time point” or “shot.” For the task given above, each run lasted 284 seconds resulting in 142 time points; eight such runs resulted in a total of 1,136 time points. The brain volume collected at each time point consisted of a stack of 30 slices; each individual slice was a 64×64 image matrix of intensity values. Each value is a measurement (see Equation 22.6) from a small $3.4 \times 3.4 \times 5$ mm box called a “voxel.” The entire volume covered a field-of-view (FoV) of $220 \times 220 \times 150$ mm. A volume was acquired slice-by-slice (67 ms per slice) over the TR, so there was a delay of almost two seconds between the time of the first slice and the last slice. See Figure 22.2 for an image of an fMRI slice (and corresponding anatomical). For the visit described above, there were a total of 122,880 ($30 \times 64 \times 64$) voxels in each time point and 1,136 time points for a total of over 139 million samples. Eighteen subjects were scanned at four sites; all subjects visited one of the sites twice making a total of five visits. For full documentation of this study, see Brown et al. (2011) and Greve et al. (2010).

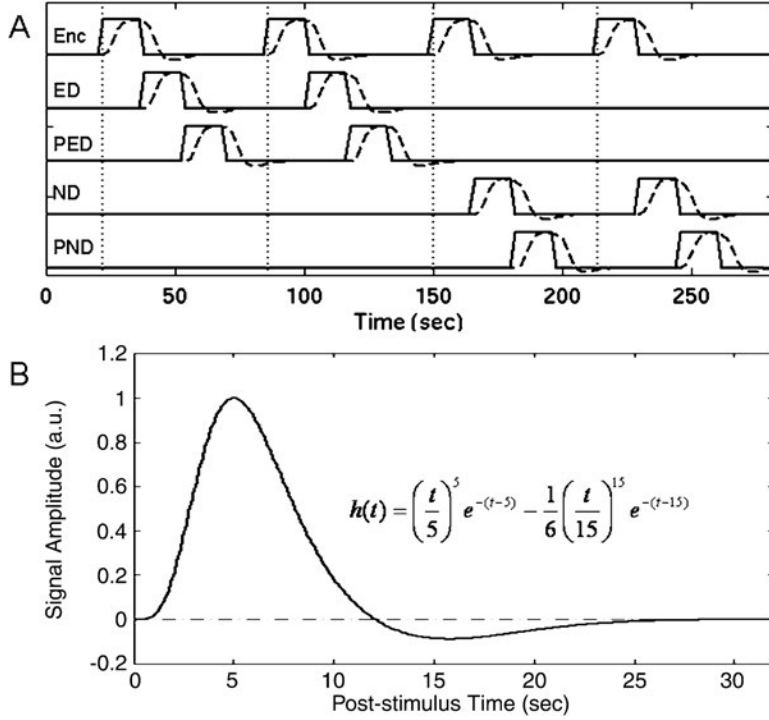


Figure 22.1 Temporal waveforms used in the analysis of the working memory task. Panel A: Task timing waveforms (solid lines) and regressors (dashed lines) for each condition for a single run. Encode (*Enc*), Emotional Distractor (*ED*), Probe following Emotional Distractor (*PED*), Neutral Distractor (*ND*), and Probe following Neutral Distractor (*PND*). The regressor is formed by convolving the assumed HRF in Panel B with the paradigm timing waveform (solid lines) to form a biologically plausible waveform (dashed lines). Panel B: Canonical shape to the HRF. This shape can be interpreted as an impulse response to a very brief stimulus. This shape and parameters are the default used by the SPM software package.

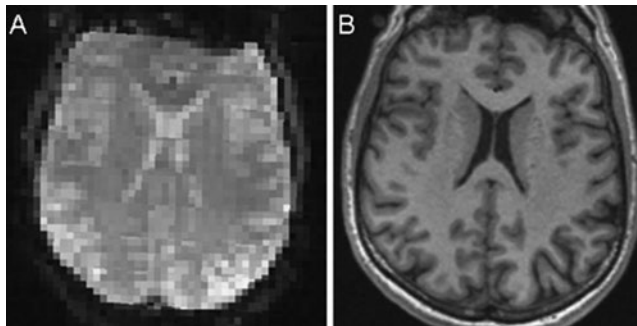


Figure 22.2 Functional (Panel A) and anatomical (Panel B) images of the same subject. The functional image has a 3.4 mm resolution and a T_2^* weighting. The anatomical image has a 1 mm voxel resolution and a T_1 weighting.

fMRI time series (first-level) analysis

While there are many ways to analyze fMRI, we will describe a typical fMRI analysis (Friston, Jezzard, & Turner, 1994; Huettel, Song, & McCarthy, 2009; Jezzard, Matthews, & Smith, 2001; Monti, 2011; Woolrich, Ripley, Brady, & Smith, 2001). First-level analysis generally consists of several steps: preprocessing, constructing a design matrix as part of a general linear model (GLM), temporal whitening, constructing contrasts to test hypotheses, computing p -values, and correcting for multiple comparisons. The preprocessing operations, like motion correction and spatial smoothing, are designed to reduce a specific kind of noise; they are discussed in detail as each source of noise is discussed below in the section on sources of time series (first-level) noise. This section describes the remaining analysis steps, how the signal is modeled and the noise is computed, and prepares the reader for tracing the propagation of noise to the higher-level analysis.

Each voxel will have its own time course (the “Raw Signal” in Figure 22.3), and the analysis is performed at each of 122,880 voxels separately by correlating its time course with a biologically plausible waveform that represents the anticipated BOLD response to the task. This waveform is constructed based on two pieces of information. First, a square waveform (solid lines in Figure 22.1, Panel A) is created using the onset and offset of the task of interest (1 for “on” and 0 for “off”). The brain’s hemodynamic response to stimulation, however, is known to be a smooth function that takes approximately four to six seconds to peak and eight to 10 seconds to return to baseline levels (Figure 22.1, Panel B), rather than the square on/off waveform of the stimulation. For this reason, the on/off pattern of stimulation is convolved with a smooth kernel like the one in Figure 22.1, Panel B to transform it into a biologically plausible waveform of up- and down-regulation of the BOLD signal (dashed lines in Figure 22.1, Panel A). The kernel is referred to as the hemodynamic response function (HRF) and can be interpreted as an impulse response. HRF shapes have been derived empirically (Friston, Frith, Turner, & Frackowiak, 1995; Glover, 1999).

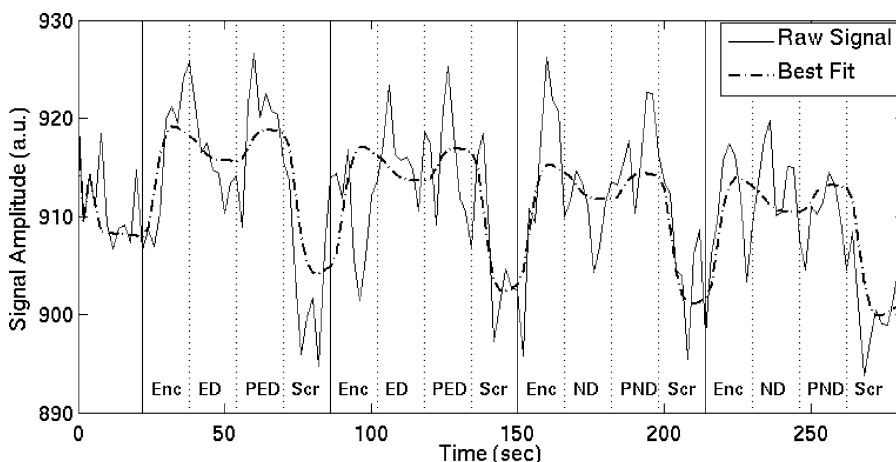


Figure 22.3 Raw and best fit time courses for a single voxel over a single run using the regressors in Figure 22.1, Panel A. Scr = Scrambled images used as baseline. Other abbreviations are the same as in Figure 22.1, Panel A.

The time course for each voxel is analyzed using a GLM (Seber & Lee, 2003):

$$Fy = FX\beta + Fn \quad (22.1)$$

where y is a vector of measurements at each time point at a single voxel (“Raw Signal” in Figure 22.3), F is a temporal filter, X is the design matrix (independent of voxel), n is the noise at the voxel, β is the unknown amplitude of the HRF at the voxel. Note that y , X , and n are vectors of length N_t (the number of time points). The columns of X (regression vectors) are formed from the biologically plausible waveforms (dashed lines in Figure 22.1, Panel A).

The quantity of interest is the amplitude of the hemodynamic response β because this relates to the neural activity at that particular location. An estimate of β can be computed using the least-mean-square (LMS) solution as

$$\hat{\beta} = (X' F' F X)^{-1} X' F' F y \quad (22.2)$$

A raw time course and resulting fit ($\hat{y} = X \hat{\beta}$) are shown in Figure 22.3.

As explained below, the fMRI noise may be temporally correlated (or “colored”), i.e., one can do better than chance at predicting what the next noise value will be given past values, which is not true for temporally uncorrelated (or “white”) noise. Temporal correlation in the noise is undesirable as it results in inaccurate computation of the p -values of the $\hat{\beta}$ estimates. Mathematically, the temporal relationship of the noise with itself is represented by the $N_t \times N_t$ temporal covariance matrix σV , where σ is the (scalar) variance of the noise and V is the matrix of auto-correlation coefficients. The value at column i , row j of V indicates the correlation between the noise waveform and itself delayed by $i-j$ time points. By definition, the value along the diagonal is 1 (i.e., $V_{ii} = 1$). White noise is indicated by a diagonal matrix ($V = I$); nonzero off-diagonal terms indicate temporally correlated noise. Temporally correlated noise can be compensated for by appropriately selecting F , the temporal filter. To remove the temporal correlation, the filter should be set such that $FV^{-1}F = I$ thereby converting the noise covariance into a diagonal matrix and producing temporal whiteness. The actual values of F can be computed as the square root of V^{-1} using the Cholesky decomposition. In practice, the temporal whitening is performed in two steps. In the first step, F is set to the identity, the GLM is solved, and an estimate of V is computed from the residuals. In the second pass, F is computed from the estimate of V , and the GLM is solved again (Burock & Dale, 2000; Woolrich et al., 2001).

If the task contains more than one condition, then X will have a column for each condition. The waveform for that condition is constructed from the part of the task waveform associated with only that condition (see Figure 22.1). Each condition will have an associated $\hat{\beta}$. In the case study above, there will be a column in X and corresponding $\hat{\beta}$ for each of the five conditions. This allows the computation of a contrast, e.g., the difference between Condition 2 (Emotional Distractor) and Condition 4 (Neutral Distractor). This contrast allows us to find brain areas that respond differentially to the emotional content of the distractor and ignore areas that respond to some other common aspect of the stimuli. This contrast is embodied in a contrast equation:

$$\hat{\gamma} = C\hat{\beta} \quad (22.3)$$

where C is the contrast matrix and $\hat{\gamma}$ is the contrast value. For the example contrast above, $C = [0 \ 1 \ 0 \ -1 \ 0]$. We can then use a t-test to test the null hypothesis that $\hat{\gamma} = 0$ with

$$t = \frac{\hat{\gamma}}{\sigma_{\gamma}}, \quad \sigma_{\gamma} = \sqrt{\frac{\sigma_r^2}{\xi}}, \quad \xi = \frac{1}{\text{trace}\left(C(X'F'FX)^{-1}C'\right)} \quad (22.4)$$

where σ_{γ} is the standard error of $\hat{\gamma}$, σ_r^2 is the variance of the residual error $e = F(\gamma - X\hat{\beta})$, and ξ is the efficiency. The significance (p -value) at the voxel can be computed from t with degrees of freedom (DOF) equal to the rows of X minus the columns of X . From this, one draws conclusions about the effect of the task at this location in the brain for this individual on this particular occasion. If the contrast is multivariate (i.e., C has more than one row), then an F -statistic would be used. The noise, summarized by its variance or standard error, is obviously very important. Below, we describe the components that go into this noise as well as efforts made to compensate for them. Note that $\hat{\gamma}$ is not an absolute quantitative measure of neural activation or energy consumption and does not carry meaningful physiological units. This means that conclusions must be statements about the relative size of responses (i.e. “this condition has a larger response than that condition”). (See Liu, Glover, Mueller, Greve, & Brown, 2012, for more information about quantification of fMRI.) Note also that the efficiency depends on the stimulus schedule and can be optimized in advance (Dale, 1999; Liu & Frank, 2004; Maus, Van Breukelen, Goebel, & Berger, 2010).

After performing this analysis at each voxel separately, the results can be displayed as an image called a statistical parametric map (SPM) as shown in Figure 22.4, Panels A–C. These images show opaque color proportional to the t -value for voxels that have a t -value greater than 3.3 (i.e., $p < .001$). This collection of supra-threshold voxels is known as “activation.” If the t -value is less

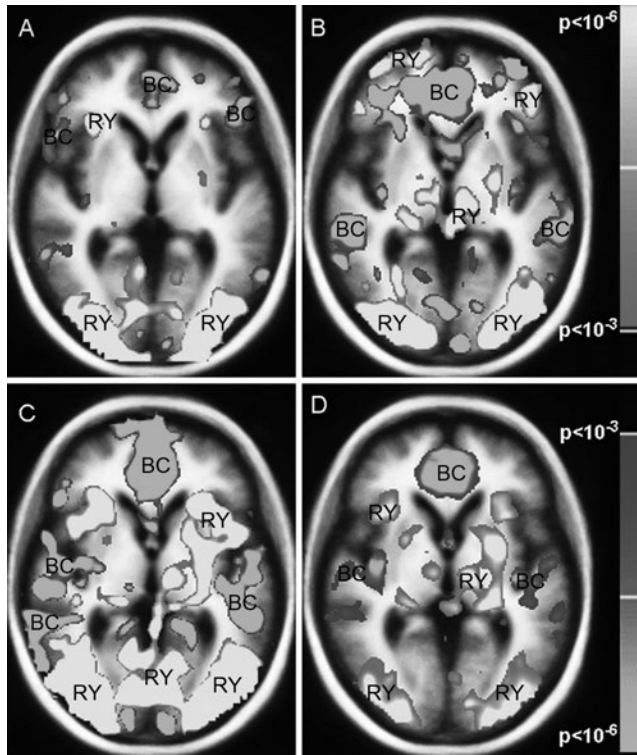


Figure 22.4 Significance SPMs for the Probe-vs.-Scrambled Contrast. Panels A–C show activation maps for three individuals. Panel D shows the results of the random-effects test over 18 subjects. Thresholds are set at $t > 3.3$ ($p < .001$, uncorrected for multiple comparisons). Red/yellow (RY) indicates Probe > Scrambled; blue/cyan (BC) indicates Scrambled > Probe.

than 3.3, then it transparently shows the gray-scale anatomical image. In this way, one can glance at an image to see if there are any significant voxels, where they are in the brain, and how much tissue they cover. The locus of activation is quantified based on the position of the supra-threshold voxels; the size of activation is quantified based on the number of supra-threshold voxels; and the intensity of activation quantified based on the $\hat{\gamma}$ in the activation region.

The presence of supra-threshold voxels needs to be interpreted with care. Each voxel represents a separate statistical test. If a p -threshold of $p < .01$ is used and there are 122,880 voxels, then one would expect 1,228 voxels to appear significant purely by chance, provided the statistical tests in each voxel are independent. In the MRI literature, this is referred to as “the problem of multiple comparisons.” One solution is to apply a Bonferroni correction by dividing the p -threshold by the number of voxels. In the above example, this would change the threshold for significance from $p < 10^{-2}$ to $p < 10^{-7}$. While this controls false positives, it dramatically increases false negatives. Another solution is to only look for “clusters” (i.e., contiguous regions of supra-threshold activity, the basic idea being that false positives will be spatially random and several adjacent supra-threshold voxels would be uncommon by chance. Closed-form distribution functions of cluster sizes have been derived under random field theory (Worsley et al., 1996); this allows a p -value to be assigned to each cluster depending on its size. These p -values can also be computed using simulations under the null hypothesis (Hayasaka & Nichols, 2003). Another alternative is to threshold using false discovery rate (FDR) in which the number of false positives with respect to the total number of positives (instead of the total number of tests) is controlled (Genovese, Lazar, & Nichols, 2002).

Higher-level analysis

In a full study, data on many (i.e., >15) subjects are typically acquired, perhaps from different groups, perhaps at different visit dates, and perhaps with some sort of intervention between visits. The analysis at the higher levels is conceptually the same as the first-level in that the lower-level GLM contrast value $\hat{\gamma}$ (and its standard error σ_{γ}) are used as input to a new GLM in which a new design matrix and contrast matrices are constructed, a new fit is performed, and a new error term is computed. The new error term will be composed of low-level noise plus new sources of noise that may be specific to the level being analyzed. In this section, we summarize how the higher-level analysis is performed and introduce the problem of anatomical variability between subjects and registration to an atlas space.

The higher-level GLM can be written as

$$Hz = H G \varphi + H u \quad (22.5)$$

where z is the vector of lower level $\hat{\gamma}$ at a given voxel, G is the group design matrix, φ is vector of regression coefficients to be estimated, u is the noise at the higher level, and H is a matrix that accounts for heteroscedasticity across the lower-level measurements. Analyses are often performed using a “random effects model” in which $H = I$ (Friston, Holmes, & Worsley, 1999). Increasingly, mixed effects models that take into account the variance from the lower level are becoming popular (Beckmann, Jenkinson, & Smith, 2003). See Zhou et al. (2013) for a thorough review of Bayesian hierarchical analysis of multilevel fMRI data, one form of mixed effects modeling. As with the first-level analysis, contrasts and p -values can be computed to create new SPMs (Figure 22.4, Panel D). At each of the higher levels, we can quantify the location, size, and amplitude of the response as described by the first level.

Prior to comparison across subjects, one needs to make sure that one is comparing the appropriate values. If two subjects each have 100,000 voxels, which voxel in Subject 1 should be

compared against which voxel in Subject 2? One solution is to align or register the brains of all subjects to an “atlas” (Collins, Neelin, Peters, & Evans, 1994). The human brain has a lot of structural variability, but there are some features that are consistent across all brains. These features are preserved in an atlas, which is typically formed from an average across a large sample of brain images from a representative population after the images have been rotated, translated, stretched, and sheared to make their features align better. More elaborate intersubject alignment schemes are becoming common (Ashburner, 2007; Fischl, Dale, Sereno, Tootell, & Rosen, 1998; Postelnicu, Zollei, & Fischl, 2009). Prior to group analysis, each subject’s brain is transformed to align its features with those of the atlas. The voxel of analysis in Equation 22.5 then corresponds to a voxel in the atlas space (also known as “standard” space). The images in Figure 22.2 are in a standard space known as “MNI152” (Collins et al., 1994).

There are several publically available software packages that will perform fMRI analysis from first level through higher levels. Some of the prominent ones are Statistical Parametric Mapping (SPM; www.fil.ion.ucl.ac.uk/spm/), FMRIB Software Library (FSL; <http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/>), and Analysis of Functional NeuroImages (AFNI; <http://afni.nimh.nih.gov/afni>).

MRI biophysics

To understand the sources of fMRI signal and noise, one needs to understand how the BOLD signal is formed. In this section, we describe the nature of the measurement, including the scanner acquisition parameters, tissue parameters, how these parameters interact to form signal, as well as scanner-related noise sources. Although a complete description of how individual nuclei interact with a magnetic field requires a quantum mechanical description, a classical physics approach is adequate to describe how groups of atoms generate the observable properties in most MR imaging (Hanson, 2008). We will focus our discussion on protons because most conventional fMRI imaging is dominated by signal from the nucleus of water-bound hydrogen atoms, which are protons, but our description is applicable to all MR observable nuclei. Our discussion is also necessarily brief; for in-depth descriptions of this material see Brown, Perthen, Liu, and Buxton (2007); Buxton (2009); Liang and Lauterbur (2000); and Nishimura (1996).

Protons exhibit a quantum mechanical property called “spin.” This spin causes the proton to have a magnetic moment and an intrinsic angular momentum. Normally, the spin axes (and corresponding magnetic moments) point in uniformly random directions canceling each other out over the volume of a voxel. When placed in the strong static magnetic field of a scanner (called the B_0 field), the distribution becomes nonuniform with slightly more spins pointing along the axis of the scanner bore. This creates a net magnetization vector in the voxel, something like a compass needle pointing north. When the net magnetic moment is perturbed from its equilibrium position using a pulse of radio frequency (RF) energy, it exhibits resonance behavior. The RF pulse causes the vector to tip at a certain angle away from the B_0 field and precess around in a circle like a spinning top that has been disturbed. When the RF pulse is turned off, it continues to move in a circle but slowly realigns with the magnetic field (a process called “relaxation”). As it relaxes, it reemits RF energy, which is then received by a coil and detected by the scanner electronics. This is the fundamental quantity measured in all MRI; all other quantities are derived from this quantity.

When the RF pulse is first applied, magnetization vectors of neighboring voxels tend to precess in phase at the resonance frequency, which is proportional to the strength of the magnetic field. However, the exact frequency of precession depends on slight perturbations the local magnetic field caused by factors such as the chemical environment and the structure of the tissue, so, over time, nearby protons within the voxel will become out of phase. The rate at which this

happens is called the T_2^* (pronounced TEE-TWO-STAR) relaxation rate. The rate at which the magnetization vector becomes realigned with the magnetic field is called the T_1 relaxation rate, which is also dependent upon the chemical/structural environment. Different tissues have different chemical/structural environments and so different T_1 and T_2^* . MRI works by creating an image where the signal intensity at each location is related to the T_1 , T_2^* , or proton (or spin) density (PD) at that location, thereby revealing the different tissue types. Figure 22.2, Panel A shows a T_2^* -weighted image; Figure 22.2 Panel B shows the same slice in a T_1 -weighted image.

An image is *not* formed by simply sampling the signal intensity at each location in the brain. When the RF pulse is transmitted, it excites the spins across a slice³ of the brain. These spins slowly reemit the excitation energy, so the measured signal is a composite from spins across the whole slice. Such a measurement actually represents a single complex point in the frequency space of the Fourier transform of the image of the slice. The frequency space is referred to as “ k -space.” Data samples are acquired at different points in k -space by making subtle manipulations to the magnetic field. When a full grid of k -space samples has been acquired, an inverse Fourier transform is applied to create an image of the slice. Typically, only the magnitude is used for fMRI. This Fourier-based image reconstruction implies that the data are sampled at known frequencies in k -space. However, small spatial variations in the B_0 field can cause this assumption to be violated, resulting in compression and stretching in parts of the image (referred to as B_0 distortion) (Jezzard & Balaban, 1995). This can also cause parts of the image to become darker than one would expect given the tissue there. Comparisons between Panels A and B in Figure 22.2 show a typical example of image distortion and signal loss by small spatial variations in the B_0 field.

The BOLD signal is created through a complex interaction between the MR physics and the living tissue. Active neurons expend energy which requires oxygen. Oxygen is delivered by hemoglobin in the blood vessels that run through the tissue. Energy usage by the tissue causes oxygen to be extracted from the blood, converting oxygenated hemoglobin (HbO) to deoxygenated hemoglobin (HbR). The BOLD signal is related to the amount of HbR present in the blood. The deoxygenated blood is pumped back to the lungs to acquire more oxygen so that the process can be repeated. The nearby vasculature react to the neural activation by dilating, producing an increase in blood flow and volume in the nearby tissue. This washes out the HbR, changing its concentration and so the BOLD signal. The key property that makes fMRI possible is that HbR has different magnetic properties than HbO. Specifically, blood with more HbR has a shorter (i.e., smaller) T_2^* . This change in blood properties causes the Blood Oxygen Level Dependent (BOLD) effect. BOLD-weighted images (e.g., Figure 22.2, Panel A) will then show the pattern of regional differences that depend upon the relative concentration of HbR, and from this the pattern of neural activity is inferred. This pattern is a snapshot because the blood flow and changing neural activation makes the pattern change over the course of a few seconds. The temporal relationship between the neural activity and the change in the BOLD signal is the hemodynamic response mentioned above (see also Figure 22.1, Panel B). This relationship is quite complicated and not yet well understood (Logothetis & Wandell, 2004). It may also depend on many physiological variables not related to neural activation (Buxton, Uludag, Dubowitz, & Liu, 2004).

The value measured at a given voxel at a given time is governed by the following equation (Liang & Lauterbur, 2000):

$$y(v, t) = \frac{PD(v) \cdot \sin(\alpha(v, t)) \cdot \left(1 - e^{\frac{-TR(v, t)}{T_1(v)}}\right) \cdot e^{\frac{-TE}{T_2^*(v, t)}}}{\left(1 - \cos(\alpha(v, t)) e^{\frac{-TR(v, t)}{T_1(v)}}\right)} + w(v, t) \quad (22.6)$$

where ν is the voxel index and t is the time of acquisition. The equation depends on three scanner acquisition parameters (TR, TE, and α) and three tissue MR parameters (T_2^* , T_1 , and PD) while $w(\nu, t)$ is thermal noise (sometimes called background noise). Acquisition parameters can be selected to emphasize certain image contrast properties (something like setting the shutter speed on a camera). The flip angle α controls how far the RF pulse tips the spins from alignment with the magnetic field. The TR (mentioned above) is the repetition time that sets the rate of RF pulses applied to each slice. Since an image can only be measured after an RF pulse, the TR effectively controls the sampling rate. The TR is typically in the range of 2–3 seconds for fMRI studies resulting in a Nyquist rate around 0.17–0.25 Hz. The TE is the echo time, which is the time delay between the RF pulse and the acquisition of the image; the TE controls the amount of T_2^* weighting. In the case study described earlier, $\alpha = 77^\circ$, TE = 30 ms, and TR = 2 seconds. The (unknown) tissue parameters indicate how the tissue interacts with the magnetic field.

An acquisition proceeds as follows: an RF pulse with flip angle α is applied to a slice of the brain, this tilts the spins by α ; after the RF pulse ends, the spins slowly realign with the main magnetic field (time constant T_1) and dephase with nearby spins (time constant T_2^*); the measured signal, γ , is dependent on the number of spins inside the voxel (PD) weighted by both the T_1 and the T_2^* decay; another RF pulse is applied after TR seconds, and the process is repeated. Imaging takes place in the scanner while the subject is engaged in a behavioral task of known timing. The value from Equation 22.6 is used as input to Equation 22.1, which is solved to compute an estimate of the HRF amplitude for an assumed HRF shape. This estimate is subject to uncertainty because of the noise in the first-level measurement.

Sources of time series (first-level) noise

Time series noise (n in Equation 22.1) manifests as any part of the measurement that does not fit the matrix model. It is summarized by its variance estimate σ_r^2 and temporal covariance V .

It creates uncertainty in the estimate of the HRF amplitude ($\hat{\gamma}$ in Equation 22.2); this uncertainty is quantified by the standard error of $\hat{\gamma}(\sigma_{\gamma}$ in Equation 22.4). Equation 22.6 links the biophysics to both the HRF amplitude and its uncertainty. Ideally, the measurement γ would only change in response to neurally driven changes in T_2^* . However, changes in any of the seven parameters in Equation 22.6 ($\alpha, \text{TR}, \text{TE}, T_1, T_2^*, \text{PD}, w$) will propagate to unmodeled changes in γ ; this includes fluctuations in T_2^* that are *not* linked to neural activation. In the remaining portion of Section 6, we trace these fluctuations first by describing scanner-related sources of noise and then move on to the more complicated subject-related sources. See Table 22.1 for a summary of sources of time series noise.

Scanner-related noise

Scanner-related noise originates both from thermal noise and from instability in the scanning process itself. The thermal noise (w in Equation 22.6) comes from Brownian motion of ions. Ions are charged particles, and the movement of charged particles is an electric current. This produces a fluctuating electromagnetic field, which is picked up by the electronics used to measure the MR signal. This electromagnetic noise is similar to the noise of a radio tuned to a frequency where there is no broadcast and is well modeled by spatially and temporally independent homoscedastic Gaussian noise.

Table 22.1 Sources of first level noise in fMRI

<i>Source</i>	<i>Mitigation</i>
Thermal	Optimize acquisition parameters Spatially smooth Eliminate external noise sources in the scanner room
Scanner instability	Diagnose and minimize with QA protocol
Head motion	Stabilize/restrain head Use prospective motion correction
Spin history	Motion-related nuisance regressors
Heartbeat	External monitor, nuisance regressors
Heart rate variability	External monitor, nuisance regressors
Respiration	External monitor, nuisance regressors
Respiration variability	External monitor, nuisance regressors
CO ₂ variability	External monitor, nuisance regressors
Neural activation	Independent component analysis
HRF model error	Include temporal derivative

Thermal noise

Thermal noise can be reduced by spatial smoothing or by adjusting the acquisition parameters. Spatial smoothing is the averaging of the waveform at a voxel with those of nearby voxels. If the signals in nearby voxels exhibit greater spatial similarity with each other than with the noise (generally true), then smoothing will result in an increase in SNR at the cost of spatial resolution. Smoothing also assures that the data meet the requirements needed by random field theory in the correction of multiple comparisons (Hayasaka & Nichols, 2003). In addition, smoothing improves intersubject alignment (discussed below). Smoothing is an extremely common pre-processing step in fMRI, and few operations will have as dramatic effect on the results as spatial smoothing (Strother et al., 2004).

The acquisition parameters (TR, TE, α) can be selected to maximize the BOLD contrast with respect to the thermal noise. While this improves the overall SNR, the effect has diminishing returns (Kruger & Glover, 2001; Triantafyllou et al., 2005). To understand why, consider that Equation 22.6 is the sum of two components. The first component is the actual MR signal (dependent upon the tissue and acquisition parameters). The second component is the thermal noise. While the thermal noise is pure noise, the MR signal has both neurally driven changes (signal in Equation 22.1) and changes due to fluctuations in the parameters that show up as noise in Equation 22.1. Thus, one can think of the BOLD measurement as having a single neurally driven signal component and two noise components, one thermal and one related to nonneural fluctuations in the MR signal. Adjusting the acquisition parameters can maximize the MR signal with respect to the thermal noise, but even if the relative thermal noise were reduced to zero, there would still be noise caused by the nonneurally driven changes in the MR signal. The actual contribution of thermal noise is hard to quantify exactly because it is highly dependent on acquisition parameters such as flip angle, TE, and voxel size. However, in the case study described in Section 2, the thermal noise generally accounted for only about 10–20% of the total temporal noise variance in gray matter (Greve et al., 2010).

Scanner instability noise

Noise caused by scanner instability can come from several sources. The RF power used to flip the spins by angle α may fluctuate slightly from shot to shot. The electronics that control the location in k -space at which a sample is made may fluctuate from shot to shot, which changes the location in k -space of a sample from the assumed frequency grid. This can cause a shot to shot change over the entire image when the inverse Fourier transform is applied. Its distribution is not well characterized, but it does have some temporal and spatial correlation. In the well-operating scanners of the case study, the instability noise only accounted for a few percent of the total variance (Greve et al., 2010). Both thermal noise and instability are present when scanning a “phantom” (i.e., an inanimate object such as a bottle of water) – so no human needs to be present to characterize these sources of image noise.

Subject motion-related noise

When a human is placed in the scanner, noise from motion and physiological effects will be introduced. An assumption of Equation 22.1 is that all the data points in the measurement vector (y) come from the same location in the brain. This location is a voxel that encompasses a tiny box inside of the scanner, and the location of this box is fixed with respect to the scanner. Thus, if the subject moves during the run, then the time course will come from one brain region during the beginning of the scan and from a different part of the brain after the motion. This can cause a large error when fitting the task time course to the voxel time course. Efforts are made to restrain the subject’s head during the MRI scan, but the head is not rigidly restrained and some motion may take place. Given that the typical fMRI voxel size is about 4 mm, even a small amount of motion can degrade the measured time course.

Motion can also interact with the MR physics in a complicated way to cause something called the “spin history effect.” This causes some parts of the brain to brighten or darken dramatically for a few time points after the motion (Friston, Williams, Howard, Frackowiak, & Turner, 1996). These effects are unmodeled in Equation 22.1 and so increase the noise. They are also irreversible.

The noise caused by motion is nonnormal, heteroscedastic, nonstationary, and temporally correlated. It can also be correlated with the task if the subject moves with the task (e.g., flinching in response to a disturbing picture). The effects of motion can be reduced by applying a motion-correction (MC) algorithm (Cox & Jesmanowicz, 1999). MC attempts to adjust the translation and rotation at each time point so that each voxel represents a single location in the brain at all time points. This cannot undo all the effects of motion (Friston et al., 1996), and itself requires some interpolation. It is worth noting that there are many efforts to measure and compensate for motion prospectively, thus reducing the effects of motion greatly (Thesen, Heid, Mueller, & Schad, 2000; Tisdall, Hess, & van der Kouwe, 2010; Ward et al., 2000).

Physiological noise

All changes in blood flow affect the BOLD signal by changing the concentration of HbR (and so the T_2^*). Ideally, the flow would only change in response to neural activation. Unfortunately, changes in flow can be caused by other factors unrelated to neural activity, such as heartbeat and respiration. These factors are referred to as “physiological noise.” The brain also has complicated autoregulatory mechanisms that attempt to keep blood flow constant (Payne, 2006). When the heart beats, it sends a surge of blood through the vasculature, and this can be detected in the BOLD signal. The heart rate is about 1 Hz, much faster than neurally driven changes in the BOLD signal. Unfortunately, the

heart rate is also faster than the Nyquist rate in most fMRI studies, resulting in the heartbeat effects being aliased into the much lower task frequencies. The heartbeat can be measured at high sample rates using an external device synchronized with the scanner. This measurement can then be used to construct “nuisance” regressors that are appended as columns to the design matrix in Equation 22.1 to model these effects (Glover, Li, & Ress, 2000). Nuisance regressors can also be constructed based on heart rate variability – low frequency changes in heartbeat that appear to account for a large fraction of the BOLD noise (Chang, Cunningham, & Glover, 2009).

Respiration can have a number of effects on the BOLD signal. As the chest cavity expands and contracts, it actually changes the magnetic field strength in the brain by tiny amounts. Though small, this causes the entire image to stretch or compress slightly with the breath (Brosch, Talavage, Ulmer, & Nyenhuis, 2002). From a noise perspective, this acts like a type of motion artifact. Respiration rates are generally about 0.3 Hz. Like heartbeat, respiration is typically faster than the BOLD Nyquist rate. The respiration rate also controls the amount of carbon dioxide (CO_2) in the blood. CO_2 causes blood vessels to dilate, which increases flow. As the respiration rate changes, the concentration of CO_2 changes. This changes the blood flow, which changes the BOLD signal. As with heartbeat, nuisance regressors can be derived from a respiration monitor and included in the BOLD model design matrix (Birn, Diamond, Smith, & Bandettini, 2006). CO_2 concentration can also be tracked directly with an external device to create nuisance regressors (Wise, Ide, Poulin, & Tracey, 2004).

The brain also has endogenous neural activity that is not related to the task (Biswal, Yetkin, Haughton, & Hyde, 1995; Chang & Glover, 2009). This activity draws blood just like task-related activity, but since there is no model for it in the design matrix the effects of this endogenous activity go into the error term. These fluctuations happen in the absence of a task so they are sometimes called “resting state networks” (RSNs). In a “functional connectivity” study, the RSNs themselves are the object of investigation (Fox et al., 2005; Vincent et al., 2006). RSNs are currently a very active area of fMRI research. The use of model-free spatio-temporal analysis has become a prevalent tool for the study of RSNs. These include the use of independent component analysis (ICA), which decomposes the entire 4D data set into independent spatial patterns and corresponding temporal waveforms (Beckmann & Smith, 2004; Calhoun, Adali, Pearlson, & Pekar, 2001). Temporal waveforms that correspond to RSN noise, as well as noise from motion and physiological sources, can be determined and removed from the analysis. Since the noise is generally low frequency, polynomial nuisance regressors are often added to the design matrix (Worsley et al., 2002).

Noise associated with HRF model errors

When analyzing the data, it is generally assumed that the neural activation follows the timing of the stimulus that the HRF takes a fixed, known shape, and that the response to each presentation is identical (e.g., as in Figure 22.1, Panel A). These assumptions can be violated in a number of ways. The shapes frequently used to model the HRF were derived from empirical data (Cohen, 1997; Friston et al., 1994; Glover, 1999), but it is well known that the shape changes across subject and brain region (Aguirre, Zarahn, & D’Esposito, 1998; Handwerker, Ollinger, & D’Esposito, 2004; Henson, Price, Rugg, Turner, & Friston, 2002; Hopfinger, Buchel, Holmes, & Friston, 2000; Miezin, Maccotta, Ollinger, Petersen, & Buckner, 2000). These HRF model errors add noise, although this is generally a small fraction of the total noise. More importantly, they bias the estimate of the response amplitude by causing the estimate to be too low. This makes the activation more difficult to detect and can cause a bias between groups (see below). The effect of these errors tends to drop with stimulus duration (Huettel et al., 2009). It is quite common to model these errors by constructing additional regressors that correspond to the temporal derivative (i.e.,

first-order Taylor series expansions) of the assumed HRF (Friston, Fletcher, Josephs, Holmes, Rugg, & Turner, 1998) or other basis sets (Woolrich, Behrens, & Smith, 2004). While this will account for some of the error, it is not sufficient by itself to recover the true amplitude of the HRF. In the first-level analysis, the additional regressors can be used in an F -test (Liu & Frank, 2004). Since only a single value is passed to higher-level analyses, these additional regressors essentially act as nuisance regressors when considered from the effect on the higher level. Bias due to small delay errors (< 1 s) can be reduced by computing a signed magnitude of the derivative and nonderivative regression coefficients (Calhoun, Stevens, Pearson, & Kiehl, 2004), though this does introduce some nonlinearity into the process. There are some nonlinear methods that attempt to fit both the shape and the amplitude (e.g., Woolrich, Jenkinson, Brady, & Smith, 2004), but these are computationally intensive and not in general use.

In addition to systematic shape errors, the responses to repeated identical stimuli may not be the same. For example, the amplitude of the response to the second of two closely spaced stimuli may be less due to refractory effects (Huettel & McCarthy, 2000). The subject may be presented with tens or hundreds of presentations of the same stimulus type (e.g., emotional distractors), and the response may change with repetition. In primary sensory areas, the response is stable across repeated presentations within visit (Miezin et al., 2000), but care must be taken to avoid practice effects (Kelly & Garavan, 2005).

Nuisance regressors

The nuisance regressors are added with the ultimate goal of reducing the uncertainty in the HRF amplitude contrast (quantified by its standard error σ_γ). However, they must be used with care. Adding regressors to the design matrix can reduce the residual variance σ_r^2 , but this can also reduce the efficiency ξ . The final effect on σ_γ will be unknown. If ξ is reduced more than the efficiency, σ_γ will decrease and detectability will increase. However, it is possible for σ_r^2 to decrease and σ_γ to increase if the efficiency is sufficiently reduced, paradoxically causing a reduction in the amount of detected activation even though more noise is accounted for.

Sources of higher-level noise

In this section, we describe the sources of noise in the higher-level analysis, where the $\hat{\gamma}$ are combined across visit, subject, and group, and perhaps site.

Within-subject/cross-visit level noise sources (Table 22.2)

Subjects are often scanned longitudinally to track disease process or evaluate an intervention. When a subject returns to be scanned for a second time, differences may appear between the

Table 22.2 Sources of within subject/cross-visit noise in fMRI

<i>Source</i>	<i>Mitigation</i>
Scanner changes	Track and reduce with QA protocol
Head placement	Use scanner auto-alignment Train technician
Subject wakefulness	Scan at same time of day Survey wakefulness
Caffeine/nicotine/etc.	Control/monitor subject chemical use

BOLD signal in the two visits that have little to do with changes in neural activity. These differences have several origins. First, the subject may not be placed in the scanner in exactly the same position, which can change the B_0 distortion. This can be reduced by using an on-line automatic slice positioning method (van der Kouwe et al., 2005) and/or by proper training of study staff to position the subject's head in the coil and position the slices. It is also possible that the scanner itself has changed between visits. This may take the form of degradation in scanner hardware over time, recalibration of the scanner, and/or hardware or software upgrades. For this reason, it is recommended that a quality assurance (QA) protocol (Friedman & Glover, 2006; Greve et al., 2010) be implemented to assure consistent scanner performance. Subjects may also have changed in ways that do not relate to the task performance but can affect the HRF amplitude. These factors include the amount of sleep the previous night (Thomas & Kwong, 2006), caffeineation level (Liu et al., 2004), and recent alcohol (Levin et al., 1998) and nicotine (Kumari et al., 2003) consumption. These can be reduced by instructing or monitoring subjects' sleep patterns and/or use of chemicals.

Within-group/cross-subject level noise sources (Table 22.3)

A group is a cohort of individuals who have been classified together by some criteria such as diagnosis (people with Alzheimer's, people with schizophrenia, healthy), handedness, gender, age, etc. These subjects will have natural variation in their HRF amplitude across the population that is purely related to functional/neural differences. However, some of the differences may be due to HRF model error mentioned above, i.e., individuals may have the same HRF amplitude but different shapes, and the different shapes cause the estimate of the HRF amplitude to differ. If the HRF error is delay-related, the effect can be reduced by adding a temporal derivative of the HRF to the design matrix then passing up to the next level a signed magnitude of the main and derivative regression coefficients as discussed above.

Anatomical variability is a major source of noise when comparing across subject. As with all analyses, the assumption is that all the values in the input data vector (z in Equation 22.5) come from the same place in the standard atlas brain, but individuals differ considerably in both brain structure and how function maps to that structure, and it is not clear exactly how this mapping between individuals should be performed. If the registration procedure does not align functional units properly, then there will be differences between the subjects because the $\hat{\gamma}$ being compared come from different brain regions. This will increase the variability within the sample, though the effect of registration errors can be reduced by spatial smoothing or by using high DOF registration methods.

Cross-group level noise sources (Table 22.3)

Comparing across groups (e.g., Alzheimer's patients versus age-matched controls) adds several more sources of noise. The anatomical variability between subjects can be even more important between groups because different groups may have systematic differences in brain structure that

Table 22.3 Sources of cross-subject and cross-group level noise in fMRI

Source	Mitigation
Anatomical variability	Use high DOF intersubject registration methods
HRF model error	Include temporal derivatives
Head motion variation across groups	Use weighted-least squares/mixed effects models

cause their registration to standard space to be systematically different (Shen, Sterr, & Szameitat, 2005). This means that at a given point in the standard brain, the HRF amplitudes from one group could come from one place in the brain while the amplitudes from a second group could come from a different place. While group correlated registration errors can lead to added variability, they can also generate a bias in the group effect. For this reason, the use of higher-order registration methods is recommended. There may also be systematic differences in the shape of the HRF between groups (D'Esposito, Zarahn, Aguirre, & Rypma, 1999), which can bias the amplitude comparisons, though these can also be reduced by passing the signed magnitude of the nonderivative and derivative components to the higher level as mentioned previously. Finally, different groups may have differences in first-level noise (Maxim et al., 2005) which can create heteroscedasticity at the higher level. This potential noise source should be controllable by the use of a mixed effects modeling in the higher-level analysis. There can also be differences in vascular physiology between groups that can affect the BOLD response. For example, the baseline cerebral blood flow (CBF) will affect the amplitude of the HRF. The CBF can change with age, the use of caffeine or other drugs, and disease state (Fleisher et al., 2009). It might be possible to reduce this confound by measuring CBF using an additional MRI acquisition called arterial spin labeling (ASL) (Alsop & Detre, 1996).

Noise sources in multiple site studies (Table 22.4)

Scanning different subjects at multiple sites can greatly increase the number and diversity of the sample pool. However, it also introduces new sources of noise related to different hardware, research protocol, and personnel (Glover et al., 2012). Scanner hardware can differ in many ways, with the two most prominent being field strength and manufacturer. Field strength directly affects the T_2^* and so systematically affects the size of the measured BOLD hemodynamic response. Even when the scanners have the same field strength, they may have been manufactured by different companies. There are details in the ways that the scanners are constructed, configured, and programmed that can cause differences in the results (Friedman et al., 2008). Many of these effects can be mitigated by appropriately selecting the scanner parameters to be as similar as possible across the scanners (Brown et al., 2011; Suckling et al., 2008). Sites may have different RF noise environments (Greve et al., 2010) due to differences in room shielding and/or the presence of electronics (such as a projector) inside the room.

Table 22.4 Sources of variation across sites in fMRI

<i>Source</i>	<i>Mitigation</i>
Field strength	Choose scanners of the same strength
Manufacturer	Analyze with B0 distortion correction To the extent possible, standardize protocol across vendors
RF noise	Use weighted-least squares/mixed effects models
Head restraint	Use same head restraint method
Stimulus delivery	Match stimulus delivery devices
Personnel/screening	Training and documentation Protocol standardization
Acquisition parameters	Match acquisition parameters as closely as possible
Head coil	Match head coils

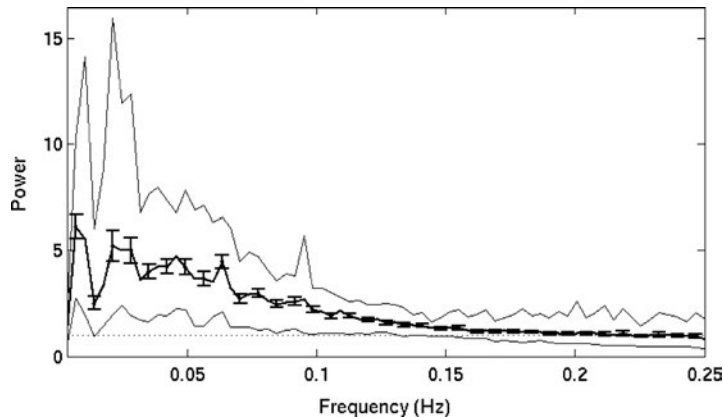


Figure 22.5 Residual noise power spectrum averaged over gray matter. The *center dark line* is the mean across 18 subjects (with *standard error bars*). The *lines above and below* are the maximum and minimum across the 18 subjects. The values have been scaled so that the mean in the high frequency (white noise) range is 1. The prominent dip in the spectrum at 0.015 Hz is due to the fitting of the task waveforms shown in Figure 1, Panel A. The time series were de-measured resulting in zero power at DC.

The subjects will likely have different experiences at different sites, and this difference may introduce systematic changes across site. For example, different sites may use different head restraints, which can affect the amount of motion and subject comfort. Sites may also differ in the way they deliver stimuli to the subject (e.g., back projection vs. goggles) or record subject responses (e.g., keyboard vs. button box). Each site will also have its own personnel responsible for executing the study. Different personnel mean that recruiting and screening practices may be different resulting in a different group sample for each site. Finally, the locale of each site may differ in terms of access to chemicals that affect the hemodynamic response (e.g., caffeine, nicotine, alcohol, etc.).

Some site differences, such as magnet vendor, may have the same differential effect on all subjects. Other site differences, such as the ease of access to nicotine or caffeine, may cause the magnitude of between-subject differences to vary by site, producing site-by-subject interactions. All of this suggests that site be taken into account when performing the higher-level analysis with data pooled from multiple sites. For a small number of sites (less than six), modeling site as a fixed effect is recommended; for a larger number, a random effects model is possible (Glover et al., 2012). It is also suggested that the experimental design be balanced across site. For a thorough review of recommendations for all aspects of multisite fMRI, see work from the Function Biomedical Informatics Research Network (Glover et al., 2012).

Discussion

Noise is important to understand and quantify because it creates uncertainty and can lead to the drawing of both false positive and false negative conclusions. With the exception of model error and possibly motion, the first-level noise sources will not bias the expected value of the estimate of the HRF amplitude. They will, of course, make it more difficult to detect activation (i.e., increased false negatives). Aside from the thermal noise, the fMRI noise is generally heteroscedastic, temporally correlated, and nonstationary (Turner & Twieg, 2005). A power spectrum of the residual time series noise from the case study described earlier is shown in Figure 22.5; the

noise is clearly nonwhite, with much power in the lower frequencies. These nonideal statistical properties bias the estimates of the standard error σ_γ , and so bias the first-level t - and F -statistics by generally making them larger than they should be (false positives) due to the low-frequency nature of the noise (Bullmore et al., 2001). This biases estimates the volume of activation computed by counting the number of voxels above threshold. The nonthermal noise generally accounts for about 80–90% of the variance in fMRI time series (Greve et al., 2010); this makes thresholded statistical maps of an individual subject highly variable (McGonigle et al., 2000). To some extent, this can be corrected by adding appropriate regressors or using of temporal whitening as discussed previously, but we recommend that the HRF amplitude estimate (or contrasts thereof) be used in higher-level analyses (Smith et al., 2005) rather than supra-threshold voxel counts or other noise-dependent statistics. The magnitude of σ_γ may be affected by visit and group membership (e.g., schizophrenics may move more than healthy controls). This can result in heteroscedasticity at higher levels. To some extent, this can be reduced by including σ_γ in a mixed effect model.

At the higher levels, the sources of noise are also very complicated with poorly understood distributions. Permutation methods (Hayasaka & Nichols, 2003) are becoming more common to control for this. At the higher level, bias is more of a concern because of systematic differences between populations that can affect the HRF amplitude estimate in ways that do not relate to underlying changes in neural activation (e.g., HRF shape variability and anatomical variability).

Methodology to reduce the impact of these sources of noise is a very active area of research. While one always wants to reduce noise, one needs to consider the cost of reduction and weigh it against the potential benefit. For example, at the first-level, the noise is dominated by physiological effects and subject motion with thermal and instability noise playing progressively smaller roles (Greve et al., 2010; Triantafyllou et al., 2005). This suggests that measures to reduce the relative contribution of thermal noise (e.g., increasing field strength) may have very little effect in the final fMRI analysis (though increasing the field strength may be very beneficial for other MRI applications). Likewise, at the level where subjects are combined, the intersubject variance may account for 50–90% of the total noise (Brown et al., 2011; Smith et al., 2005; Suckling et al., 2008), depending upon the number of stimulus presentations and other factors. Thus, efforts to reduce first-level noise of any kind, even when successful, might have very little impact at the higher levels.

Conclusion

The purpose of fMRI is to draw conclusions about neural activation in an individual at a single time point or across time points, across individuals, or across groups. Noise is introduced at each of these levels and causes uncertainty and/or bias in the final conclusions. At the lowest level, fMRI reveals changes in neural activation based on changes in the deoxygenation of blood (the BOLD effect). The analysis is performed by fitting the (known) stimulation time course with the BOLD signal at each point in the brain. This yields an estimate of the amplitude of the BOLD signal in response to the stimulus. This amplitude is then used as a surrogate of neural activation, though the relationship between the two is complicated and not well understood. Noise at this level manifests itself as error in the fit. This noise may be due to the scanner (thermal noise or instability), subject motion, deviations between the assumed and actual HRF shape, or any physiological effect that changes blood flow, but is not related to neural activation (e.g., respiration). Across days or weeks, variation can be caused by changes in the scanner, subject wakefulness level, and consumption of vaso-active chemicals such as caffeine. When comparing across subjects, the subjects' brains must be registered to a common space. Errors in this registration can cause error

at the higher-level analysis. Group differences in the underlying shape of the HRF can create differences in the higher-level analysis even when no amplitude difference exists. In a multisite study, vendor and field-strength differences can also be a source of noise.

The field of fMRI is very rich in methodology, both in terms of acquisition and analysis, and this survey has only touched on the most general of these methods. fMRI data is also rich in noise sources, not all of which have been reviewed here, and some of which are still being discovered and understood. While noise in fMRI does represent a substantial challenge to its practical use, fMRI has been successfully used in a large number of studies to map the functions of the human brain.

Acknowledgements

Support for this research was provided in part by the National Center for Research Resources (P41-RR14075, R01 RR16594, P41-009874, the NCRR BIRN Morphometric Project BIRN002, and Functional Imaging Biomedical Informatics Research Network (FBIRN) U24 RR021382), the National Institute for Biomedical Imaging and Bioengineering (R01 EB001550, R01EB006758), as well as by the Department of Energy (DE-F02-99ER62764-A012) to the Mind Research Network (previously known as the MIND Institute).

Notes

- 1 Originally published as: Greve, D.N., Brown, G.G., Mueller, B.A., Glover, G., & Liu, T.T. (2012). A survey of the sources of noise in FMRI. *Psychometrika*, 78, 396–416. © 2012. The Psychometric Society. Used with permission.
- 2 Increasingly fMRI experiments are being performed without a task. These techniques are not extensively described in this paper as their methodology strongly overlaps with task-based analysis.
- 3 Prior to RF transmission, a third set of coils creates a controlled spatial gradient in the B_0 field so that only the spins in a given slice are on-resonance. The location of this slice is adjusted each shot so that all slices in the brain are imaged. Other types of MRI acquisitions will excite the spins across the entire brain.

References

- Aguirre, G.K., Zarahn, E., & D'Esposito, M. (1998). The variability of human, BOLD hemodynamic responses. *NeuroImage*, 8, 360–369.
- Alsop, D. C., & Detre, J.A. (1996). Reduced transit-time sensitivity in noninvasive magnetic resonance imaging of human cerebral blood flow. *Journal of Cerebral Blood Flow and Metabolism*, 16, 1236–1249.
- Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, 38, 95–113.
- Beckmann, C., Jenkinson, M., & Smith, S. (2003). General multi-level linear modelling for group analysis in FMRI. *NeuroImage*, 20, 1052–1063.
- Beckmann, C.F., & Smith, S.M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23, 137–152.
- Birn, R.M., Diamond, J.B., Smith, M.A., & Bandettini, P.A. (2006). Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. *NeuroImage*, 31, 1536–1548.
- Biswal, B., Yetkin, F., Haughton, V., & Hyde, J. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar imaging. *Magnetic Resonance in Medicine*, 34, 537–541.
- Brosch, J.R., Talavage, T.M., Ulmer, J.L., & Nyenhuis, J.A. (2002). Simulation of human respiration in fMRI with a mechanical model. *IEEE Transactions on Biomedical Engineering*, 49, 700–707.
- Brown, G.G., Mathalon, D.H., Stern, H., Ford, J., Mueller, B., Greve, D.N., . . . Potkin, S.G. (2011). Multisite reliability of cognitive BOLD data. *NeuroImage*, 54, 2163–2175.
- Brown, G.G., Perthen, J.E., Liu, T.T., & Buxton, R.B. (2007). A primer on functional magnetic resonance imaging. *Neuropsychology Review*, 17, 107–125.

- Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., . . . Brammer, M. (2001). Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. *Human Brain Mapping, 12*(2), 61–78.
- Burock, M. A., & Dale, A.M. (2000). Estimation and detection of event-related fMRI signals with temporally correlated noise: a statistically efficient and unbiased approach. *Human Brain Mapping, 11*, 249–260.
- Buxton, R. B. (2009). *Introduction to functional magnetic resonance imaging: Principles and techniques* (2nd ed.). New York, NY: Cambridge University Press.
- Buxton, R. B., Uludag, K., Dubowitz, D. J., & Liu, T. T. (2004). Modeling the hemodynamic response to brain activation. *NeuroImage, 23*(Suppl. 1), S220–S233.
- Calhoun, V.D., Adali, T., Pearlson, G.D., & Pekar, J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping, 14*, 140–151.
- Calhoun, V.D., Stevens, M. C., Pearlson, G.D., & Kiehl, K. A. (2004). fMRI analysis with the general linear model: Removal of latency-induced amplitude bias by incorporation of hemodynamic derivative terms. *NeuroImage, 22*, 252–257.
- Chang, C., Cunningham, J. P., & Glover, G. H. (2009). Influence of heart rate on the BOLD signal: The cardiac response function. *NeuroImage, 44*, 857–869.
- Chang, C., & Glover, G. H. (2009). Effects of model-based physiological noise correction on default mode network anti-correlations and correlations. *NeuroImage, 47*, 1448–1459.
- Cohen, M. S. (1997). Parametric analysis of fMRI data using linear systems methods. *NeuroImage, 6*, 93–103.
- Collins, D. L., Neelin, P., Peters, T. M., & Evans, A. C. (1994). Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of Computer Assisted Tomography, 18*, 192–205.
- Cox, R., & Jesmanowicz, A. (1999). Real-time 3D image registration for functional MRI. *Magnetic Resonance Imaging, 42*, 1014–1018.
- D'Esposito, M., Zarahn, E., Aguirre, G. K., & Rypma, B. (1999). The effect of normal aging on the coupling of neural activity to the bold hemodynamic response. *NeuroImage, 10*, 6–14.
- Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping, 8*, 109–114.
- Fischl, B., Dale, A. M., Sereno, M. I., Tootell, R.B.H., & Rosen, B. R. (1998). A coordinate system for the cortical surface. *NeuroImage, 7*(4), S740.
- Fleisher, A. S., Podraza, K. M., Bangen, K. J., Taylor, C., Sherzai, A., Sidhar, K., . . . Buxton, R. B. (2009). Cerebral perfusion and oxygenation differences in Alzheimer's disease risk. *Neurobiology of Aging, 30*, 1737–1748.
- Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America, 102*(27), 9673–9678.
- Friedman, L., & Glover, G. H. (2006). Report on a multicenter fMRI quality assurance protocol. *Journal of Magnetic Resonance Imaging, 23*, 827–839.
- Friedman, L., Stern, H., Brown, G. G., Mathalon, D. H., Turner, J., Glover, G. H., . . . Potkin, S. G. (2008). Test-retest and between-site reliability in a multicenter fMRI study. *Human Brain Mapping, 29*, 958–972.
- Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., & Turner, R. (1998). Event-related fMRI: characterizing differential responses. *NeuroImage, 7*, 30–40.
- Friston, K. J., Frith, C. D., Turner, R., & Frackowiak, R. S. (1995). Characterizing evoked hemodynamics with fMRI. *NeuroImage, 2*, 157–165.
- Friston, K. J., Holmes, A. P., & Worsley, K. J. (1999). How many subjects constitute a study? *NeuroImage, 10*(1), 1–5.
- Friston, K. J., Jezzard, P., & Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping, 1*, 153–171.
- Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., & Turner, R. (1996). Movement-related effects in fMRI time-series. *Magnetic Resonance in Medicine, 35*, 346–355.
- Genovese, C. R., Lazar, N. A., & Nichols, T. E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage, 15*, 870–878.
- Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage, 9*, 416–429.
- Glover, G. H., Li, T. Q., & Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magnetic Resonance in Medicine, 44*, 162–167.
- Glover, G. H., Mueller, B. A., Turner, J. A., van Erp, T. G., Liu, T. T., Greve, D. N., . . . Potkin, S. G. (2012). Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. *Journal of Magnetic Resonance Imaging, 36*, 39–54.

- Greve, D. N., Mueller, B. A., Liu, T., Turner, J. A., Voyvodic, J., Yetter, E., . . . Glover, G. (2010). A novel method for quantifying scanner instability in fMRI. *Magnetic Resonance in Medicine*, 65, 1053–1061.
- Handwerker, D. A., Ollinger, J. M., & D’Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, 21, 1639–1651.
- Hanson, L. (2008). Is quantum mechanics necessary for understanding magnetic resonance? *Concepts in Magnetic Resonance. Part A* 32A, 329–340.
- Hayasaka, S., & Nichols, T. E. (2003). Validating cluster size inference: Random field and permutation methods. *NeuroImage*, 20, 2343–2356.
- Henson, R. N., Price, C. J., Rugg, M. D., Turner, R., & Friston, K. J. (2002). Detecting latency differences in event-related BOLD responses: Application to words versus nonwords and initial versus repeated face presentations. *NeuroImage*, 15, 83–97.
- Hopfinger, J. B., Buchel, C., Holmes, A. P., & Friston, K. J. (2000). A study of analysis parameters that influence the sensitivity of event-related fMRI analyses. *NeuroImage*, 11, 326–333.
- Huettel, S. A., & McCarthy, G. (2000). Evidence for a refractory period in the hemodynamic response to visual stimuli as measured by MRI. *NeuroImage*, 11, 547–553.
- Huettel, S. A., Song, A. W., & McCarthy, G. (2009). *Functional magnetic resonance imaging* (2nd ed.). Sunderland, MA: Sinauer Associates.
- Jezzard, P., & Balaban, R. S. (1995). Correction for geometric distortion in echo planar images from B_0 field variations. *Magnetic Resonance in Medicine*, 34, 65–73.
- Jezzard, P., Matthews, P., & Smith, S. M. (2001). *Functional MRI: an introduction to methods*. Oxford: Oxford University Press.
- Kelly, A. M., & Garavan, H. (2005). Human functional neuroimaging of brain changes associated with practice. *Cerebral Cortex*, 15, 1089–1102.
- Kruger, G., & Glover, G. H. (2001). Physiological noise in oxygenation-sensitive magnetic resonance imaging. *Magnetic Resonance in Medicine*, 46, 631–637.
- Kumari, V., Gray, J. A., ffytche, D. H., Mitterschiffthaler, M. T., Das, M., Zachariah, E., . . . Sharma, T. (2003). Cognitive effects of nicotine in humans: An fMRI study. *NeuroImage*, 19, 1002–1013.
- Levin, J. M., Ross, M. H., Mendelson, J. H., Kaufman, M. J., Lange, N., Maas, L. C., . . . Renshaw, P. F. (1998). Reduction in BOLD fMRI response to primary visual stimulation following alcohol ingestion. *Psychiatry Research*, 82, 135–146.
- Liang, Z., & Lauterbur, P. C. (2000). *Principles of magnetic resonance imaging: A signal processing perspective*. Piscataway, NJ: IEEE Press.
- Liu, T. T., Behzadi, Y., Restom, K., Uludag, K., Lu, K., Buracas, G. T., . . . Buxton, R. B. (2004). Caffeine alters the temporal dynamics of the visual BOLD response. *NeuroImage*, 23, 1402–1413.
- Liu, T. T., & Frank, L. R. (2004). Efficiency, power, and entropy in event-related FMRI with multiple trial types. Part I: theory. *NeuroImage*, 21, 387–400.
- Liu, T. T., Glover, G. H., Mueller, B. A., Greve, D. N., & Brown, G. G. (2012). An introduction to normalization and calibration methods in functional MRI. *Psychometrika*, 78, 308–321.
- Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the BOLD signal. *Annual Review of Physiology*, 66, 735–769.
- Maus, B., Van Breukelen, G. J., Goebel, R., & Berger, M. P. (2010). Optimization of blocked designs in fMRI studies. *Psychometrika*, 56, 1338–1352.
- Maxim, V., Sendur, L., Fadili, J., Suckling, J., Gould, R., Howard, R., & Bullmore, E. (2005). Fractional Gaussian noise, functional MRI and Alzheimer’s disease. *NeuroImage*, 25, 141–158.
- McGonigle, D. J., Howseman, A. M., Athwal, B. S., Friston, K. J., Frackowiak, R. S., & Holmes, A. P. (2000). Variability in fMRI: An examination of intersession differences. *NeuroImage*, 11(6 Pt. 1), 708–734.
- Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E., & Buckner, R. L. (2000). Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage*, 11(6 Pt. 1), 735–759.
- Monti, M. M. (2011). Statistical analysis of fMRI time-series: A critical review of the GLM approach. *Frontiers in Human Neuroscience*, 5, 28.
- Nishimura, D. G. (1996). *Principles of magnetic resonance imaging*. Stanford, CA: Stanford University.
- Payne, S. J. (2006). A model of the interaction between autoregulation and neural activation in the brain. *Mathematical Biosciences*, 204, 260–281.
- Postelnicu, G., Zollei, L., & Fischl, B. (2009). Combined volumetric and surface registration. *IEEE Transactions on Medical Imaging*, 28, 508–522.
- Seber, G. A. F., & Lee, A. J. (2003). *Linear regression analysis* (2nd ed.). Hoboken, NJ: Wiley.

- Shen, S., Sterr, A., & Szameitat, A. (2005). A template effect study on voxel-based morphometry in statistic parametric mapping. *Conference Proceedings (IEEE Engineering in Medicine and Biology Society Conference)*, 3, 3051–3054.
- Smith, S. M., Beckmann, C. F., Ramnani, N., Woolrich, M. W., Bannister, P. R., Jenkinson, M., . . . McGo-nigle, D. J. (2005). Variability in fMRI: A re-examination of inter-session differences. *Human Brain Mapping*, 24, 248–257.
- Strother, S., La Conte, S., Hansen, L. K., Anderson, J., Zhang, J., Pulapura, S., & Rottenberg, D. A. (2004). Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance met-rics: I. A preliminary group analysis. *NeuroImage*, 23, S196–S207.
- Suckling, J., Ohlssen, D., Andrew, C., Johnson, G., Williams, S. C., Graves, M., . . . Bullmore, E. (2008). Components of variance in a multicentre functional MRI study and implications for calculation of sta-tistical power. *Human Brain Mapping*, 29, 1111–1122.
- Thesen, S., Heid, O., Mueller, E., & Schad, L. R. (2000). Prospective acquisition correction for head motion with image-based tracking for real-time fMRI. *Magnetic Resonance in Medicine*, 44, 457–465.
- Thomas, R. J., & Kwong, K. (2006). Modafinil activates cortical and subcortical sites in the sleep-deprived state. *Sleep*, 29, 1471–1481.
- Tisdall, M., Hess, A., & van der Kouwe, A. (2010, February). *Prospective motion correction of anatomical brain sequences using EPI navigators*. Paper presented at the ISMRM Workshop on Current Concepts of Motion Correction for MRI & MRS, Kitzbühel, Tyrol, Austria.
- Triantafyllou, C., Hoge, R. D., Krueger, G., Wiggins, C. J., Potthast, A., Wiggins, G. C., & Wald, L. L. (2005). Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition param-eters. *NeuroImage*, 26, 243–250.
- Turner, G. H., & Twieg, D. B. (2005). Study of temporal stationarity and spatial consistency of fMRI noise using independent component analysis. *IEEE Transactions on Medical Imaging*, 24, 712–718.
- van der Kouwe, A. J., Benner, T., Fischl, B., Schmitt, F., Salat, D. H., Harder, M., . . . Dale, A. M. (2005). On-line automatic slice positioning for brain MR imaging. *NeuroImage*, 27, 222–230.
- Vincent, J. L., Snyder, A. Z., Fox, M. D., Shannon, B. J., Andrews, J. R., Raichle, M. E., & Buckner, R. L. (2006). Coherent spontaneous activity identifies a hippocampal-parietal memory network. *Journal of Neurophysiology*, 96, 3517–3531.
- Ward, H. A., Riederer, S. J., Grimm, R. C., Ehman, R. L., Felmlee, J. P., & Jack, C. R., Jr. (2000). Prospective multiaxial motion correction for fMRI. *Magnetic Resonance in Medicine*, 43, 459–469.
- Wise, R. G., Ide, K., Poulin, M. J., & Tracey, I. (2004). Resting fluctuations in arterial carbon dioxide induce significant low frequency variations in BOLD signal. *NeuroImage*, 21, 1652–1664.
- Woolrich, M., Ripley, B., Brady, J., & Smith, S. (2001). Temporal autocorrelation in univariate linear mod-elling of FMRI data. *NeuroImage*, 14, 1370–1386.
- Woolrich, M. W., Behrens, T. E., & Smith, S. M. (2004). Constrained linear basis sets for HRF modelling using variational Bayes. *NeuroImage*, 21, 1748–1761.
- Woolrich, M. W., Jenkinson, M., Brady, J. M., & Smith, S. M. (2004). Fully Bayesian spatio-temporal mod-eling of FMRI data. *IEEE Transactions on Medical Imaging*, 23, 213–231.
- Worsley, K. J., Liao, C. H., Aston, J., Pere, V., Duncan, G. H., Morales, E., & Evans, A. C. (2002). A general statistical analysis for fMRI data. *NeuroImage*, 15, 1–15.
- Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4(1), 58–73.
- Zhou, B., Konstorum, A., Duong, T., Tiue, K. H., Wells, W. M., Brown, G. G., . . . Shahbaba, B. (2013). A hierarchical modeling approach to data analysis and study design in a multi-site experimental fMRI study. *Psychometrika*, 78, 260–278.

This page intentionally left blank

PART VI

Special topics

This page intentionally left blank

COMPARATIVE EFFECTIVENESS RESEARCH AND META-ANALYSIS

Nicky Cullum and Jo Dumville

Comparative effectiveness research (CER) is research that compares the results of one approach for managing a disease to the results of other approaches (Agency for Healthcare Research and Quality [AHRQ], n.d.). In a nutshell, CER is concerned with finding out “what works” in health care. Decision makers use high quality, comprehensive findings about benefits and harms associated with a particular course of action to inform decisions about clinical care and health policy. The CER approach can be applied to any aspect of health care whether surgery, configuration of services, or methods of pain relief. This chapter is concerned with how the available evidence across a particular health care topic is summarized in systematic reviews and how, within a systematic review, results of separate studies are statistically pooled using meta-analysis.

Comparative effectiveness is concerned primarily with cause and effect relationships: does a given intervention *reduce* specific symptoms, or *increase* specified healthy outcomes? (For discussion of cause and effect, see Eckardt & Rindskopf, this volume; Rubin & Zell, this volume.) Findings from randomized controlled trials (RCTs) and quasi-experimental (observational) studies that compare benefits and/or harms of alternative interventions for a given health problem constitute the evidence base or building blocks for judging comparative effectiveness, as shown in Figure 23.1. The best decisions about comparative effectiveness are made when *all* findings from relevant primary research studies are identified, collated, synthesized, and analyzed using the scientific method of systematic review.

Historical developments

Summarizing evidence across studies

It seems astonishing now to remember how it was once perfectly acceptable for leading clinicians to write textbooks and review articles in which they merely summarized the research studies they knew of, or agreed with, and recommended practice based on them. We make no apology here for revisiting the best advertisement for systematic reviews ever written. It comes from medicine but there is no better example. In the 1990s, Elliott Antman and colleagues assembled all the RCTs of interventions for reducing the risk of myocardial infarction (Antman, Lau, Kupelnick, Mosteller, & Chalmers, 1992). They then undertook a cumulative meta-analysis of these studies. Results published from 1960 to 1988 were pooled in chronological order, with findings added

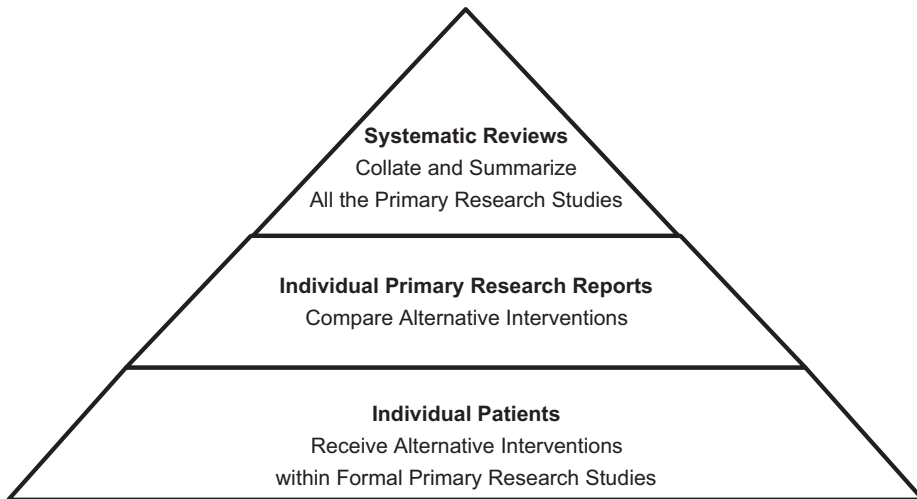


Figure 23.1 Primary research participants, studies, and systematic reviews.

one by one, sequentially and statistically. The 182 RCTs, 70 of which evaluated thrombolytics, yielded a pooled sample size of 48,154. The advantage of conducting this research on a topic like myocardial infarction is the existence of a large body of RCTs, and because they are drug studies of reasonable quality and size (since this type of evidence is required for drug licensing). The genius of this piece of work lay in the next step, which was to compare recommendations for the treatment of myocardial infarction made by experts in review articles and textbook chapters with what the accumulated evidence was actually telling us contemporaneously. The researchers classified expert recommendations as “routine” (use routinely in the absence of contraindications), “specific” (use in selected patients with particular indications), “rare/never” (use only rarely or never), “experimental” (use only within research), and “not mentioned.” They showed that whilst sufficient evidence that thrombolysis saves lives had accumulated by 1973 (at which time 10 studies involving 2,544 people had been published), experts were not recommending thrombolytics even for specific indications until 13 years later. It is sobering to consider the thousands of deaths that could have been avoided throughout the world in the interim. Moreover, once a statistically significant treatment effect had been shown in 1973, the size of the treatment effect did not change much with more research, but the precision of the estimate was increased as a further 45,000 people were recruited into (unnecessary) RCTs.

The Antman et al. (1992) paper demonstrates two essential principles for CER and evidence based practice:

Guidance or recommendations for policy and practice should be based on a comprehensive and rigorous synthesis of all the relevant research that addresses the topic.

If this rule is not followed, it is highly likely that patients will be denied effective interventions or given ineffective or even harmful ones.

New primary research should only be undertaken once there has been a comprehensive and rigorous synthesis of all the relevant research to identify real and important gaps in our knowledge.

Without such reviews as a basis for the research agenda, unnecessary research on questions that have already been answered will be conducted, important lessons from existing research will not be learned in a timely fashion, and/or the wrong research questions may receive priority.

Emergence of systematic reviews

A further concern was that most review articles, even in top medical journals in the 1980s, were vague in purpose, nonsystematic in selection of literature, qualitative in approach to synthesis of findings across studies, and nonreplicable (Mulrow, 1987). Nowadays it would be almost unheard of to find clinical practice guidelines that were not based on one or more systematic reviews, which are expected as the evidentiary basis for clinical practice guidelines (Institute of Medicine, 2011; Sabharwal et al., 2013).

The methodologies of systematic review and meta-analysis have their roots in social science and educational research (e.g., Glass, 1977). Health scientists (including nurse scientists) were early adopters. The earliest indexed meta-analysis in PubMed, co-authored by Glass, concerns the effects of psychotherapy compared with no treatment on various outcome measures (Smith & Glass, 1977). Papers about meta-analysis methods appeared in nursing journals starting in the early 1980s (O'Flynn, 1982). An early review of systematic reviews in nursing identified 36 that tackled comparative effectiveness questions; of these, 19 (53%) met the three quality criteria of (a) a clear question; (b) a comprehensive search strategy; and (c) appropriate data synthesis (Droogan & Cullum, 1998). The year 1993 saw the birth of the Cochrane Collaboration (2013), an international endeavor that seeks to prepare, maintain, and disseminate high-quality systematic reviews about the effects of health care, and in which nurses were early participants.

Cochrane Collaboration

The germ of the idea for the Cochrane Collaboration came from the writings of epidemiologist Archie Cochrane (Cochrane Collaboration, 2013), who is particularly remembered for his advocacy of the RCT as a tool to distinguish effective from ineffective health care (Hill, 2000). Cochrane (1979, p. 9) had written “*It is surely a great criticism of our profession that we have not organized a critical summary by specialty or subspecialty, up-dated periodically, of all relevant randomized controlled trials.*” In response, Sir Iain Chalmers and colleagues (1986; Dickersen, 1986) established a database of trials in perinatal care and then conducted systematic reviews of these trials. The appeal of this rigorous approach to summarizing evidence was obvious and immediate and in 1992 the UK Cochrane Centre opened in Oxford. Funded by the UK National Health Service Research and Development Programme, its brief was “to facilitate and co-ordinate the preparation and maintenance of systematic reviews of randomized controlled trials of health care.” The following year this activity mushroomed into the international Cochrane Collaboration. The scale and productivity of the Cochrane Collaboration increased dramatically; as of May 2015, there were 52 collaborative review groups (www.cochrane.org/contact/review-groups), each taking responsibility for producing and maintaining high-quality systematic reviews in a focused topical area. The topics are multidisciplinary. Review teams and editors usually come from a health profession plus epidemiology, biostatistics, and the social sciences. Most review groups cover topics that are of interest to nurses; e.g., the Wounds Group (www.wounds.cochrane.org), the Stroke Group (www.stroke.cochrane.org), and the Effective Practice and Organisation of Care Group (www.epoc.cochrane.org). Nurses are encouraged and welcomed as review group editors, lead and co-reviewers. Cochrane systematic reviews are published in the Cochrane Database

of Systematic Reviews (within the Cochrane Library), are distinguished by their high quality (Moseley, Elkins, Herbert, Maher, & Sherrington, 2009; Olsen et al., 2001), and are updated as new, relevant primary research findings are published.

Conducting systematic reviews

The sine qua non of a systematic review is that it is scientifically rigorous. Systematic reviews offer a comprehensive and unbiased answer to a specific question. The review is completed using a standard sequence of steps that are anticipated in a written protocol.

Developing a focused, answerable question

Literature can be reviewed in many ways (e.g., scholarly review, literature review, scoping review, or narrative review). Integrative reviews are written for the purpose of symbolic or conceptual utilization of findings. Integrative reviews have titles such as: “End of life care education: An in-depth review” or “Applications of focus group methodology in nursing research.” Integrative reviews give the author room for creativity and are not constrained by the scientific method, precisely because they are not seeking to answer a focused question.

This chapter is concerned with systematic review, not integrative review. Systematic review is essential when studying interventions and outcomes, or whenever findings from primary studies are used to make decisions about some actions in the process of providing care. Systematic review is used to guide practice actions whenever there is a danger of drawing misleading conclusions and misdirecting care.

The distinguishing feature of a systematic review is use of the scientific method to answer a *focused* question; a question that conceivably has an answer. The most straightforward kind of question posed for systematic review has three possible answers: yes, no, or unclear. Such questions are typically posed in Patient-Intervention-Comparator-Outcome format (PICO; Fineout-Overholt & Johnston, 2005). Examples include: Does lying young babies prone (compared with supine) reduce sudden infant death events? Does larval therapy (compared with no larval therapy) reduce the time to healing of sloughy venous leg ulcers? Does a problem-based learning approach to student nurse education improve learning outcomes compared with a traditional didactic approach? Examples of less focused, though still potentially answerable questions are: What factors increase the risk of sudden infant death in young babies? Which topical treatments reduce the time to healing of sloughy venous leg ulcers? What is the most effective teaching style for undergraduate nurse education?

Topic selection for most systematic reviews involves a trade-off between breadth and feasibility. More focused questions are easier to answer because the volume of relevant literature will be smaller. Answering broader questions is often labor intensive as the scope is much wider; however, these reviews may be very useful to decision makers because many options for course of action are considered.

The PICO method (Fineout-Overholt & Johnston, 2005) is a useful device for naming parameters of interest and framing uncertainty. *P* represents the patient population of interest, *I* is the intervention of interest, *C* is the comparator of interest, and *O* is the outcome of interest. Each parameter must be defined in advance. This focused question will then drive all other aspects of the systematic review methodology including search strategy, nature of the eligible studies, and analytical approach. Detailed operational definitions are required for each PICO element. The operational definitions are developed and refined during protocol development.

Bias

Systematic reviews are reports of a special type of observational study where primary research studies are the observed “cases.” Bias is a risk unless strenuous steps are taken to avoid it. In CER, bias refers to any factor, recognized or not, that distorts the findings of a study (AHRQ, n.d.). Within individual RCTs, bias assessment focuses on five domains: selection bias (sequence generation and allocation concealment); performance bias (blinding of participants and personnel); detection bias (blinding of outcome assessment); attrition bias (completeness of outcome data); and reporting bias (selective reporting) (Higgins & Green, 2011). Within the literature, publication bias (the greater tendency for studies with significant findings to be published) and reporting bias (time lags, duplicate publication, language, database index, media attention) can distort systematic reviews (see Box 1 in Song, Eastwood, Gilbody, Duley, & Sutton, 2000).

Steps

Generally, systematic reviews are conducted by following these steps sequentially: (a) developing a focused, answerable question; (b) publishing a protocol for the review and having it peer reviewed; (c) prospectively registering the proposed systematic review in an international register; (d) searching for eligible research studies; (e) selecting eligible research studies; (f) extracting important information from included studies; (g) analyzing data; (h) interpreting findings; and (i) disseminating results. We summarize each step; however, more detailed instruction should be sought when planning to undertake a systematic review. We particularly recommend the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins & Green, 2011; www.cochrane.org/training/cochrane-handbook); *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* (“Methods Guide,” 2014); and guidance from the Centre for Reviews and Dissemination in York, UK (Centre for Reviews and Dissemination [CRD], University of York, 2009; www.york.ac.uk/inst/crd/SysRev/!SSL/!WebHelp/SysRev3.htm).

Protocol for the review

As with primary research, a key step in reducing bias is to make detailed plans in advance, committing to paper the results of as many decisions as possible prior to seeing the data. Review protocols help to ensure transparency and rigor. Review protocols are detailed plans written in advance that (a) provide rationale for the review, (b) state objectives in PICO format, (c) describe unambiguous criteria for inclusion and exclusion of studies, (d) describe the search strategy and decision rules for applying inclusion/exclusion criteria to select papers discovered during the search, (e) identify which data will be extracted from each study, and (f) define the approach to data analysis. Protocols should be peer reviewed and once set, systematic review protocols should be registered.

Rationale

Provide a rationale for your topic of interest. The background and significance should include relevant basic epidemiology, consequences of the health issue, alternative treatment strategies, current guidelines, evidence of variations in practice, and existence of past systematic reviews and whether they are up-to-date.

Objectives in PICO format

Clearly state the question or objective in PICO format. An example of a well-framed systematic review objective is: *To compare the effects of wearing compression stockings versus not wearing compression stockings on the rate of deep vein thrombosis in passengers on long haul flights*. All PICO elements are specified in the question including the primary outcome of deep vein thrombosis. The primary outcome should be a health status measure or a clearly defined event that is important and meaningful to those most affected (e.g., patients, decision makers, the public). An outcome should not be selected simply because it may have been most reported in studies. More than one primary outcome may be identified (e.g., one benefit and one adverse effect) but to maintain focus, there should be as few as possible.

Unambiguous eligibility criteria

Eligibility criteria for studies to be included should be stated unambiguously. These will typically follow from the PICO parameters and also include design of the primary study, language of publication, and publication status. All relevant options should be considered and ambiguities should be anticipated so that inclusion/exclusion decisions can be based on detailed criteria specified in advance of seeing the studies. These steps reduce the risk that study inclusion decisions are influenced by characteristics such as research setting, results, the investigators, the source of funding. Table 23.1 highlights these issues using the flight stockings example.

Table 23.1 Study eligibility considerations: protocol for systematic review. Protocol for systematic review of effects of compression stockings for DVT prevention in long-haul flight passengers

<i>Characteristic</i>	<i>Possible Criterion</i>	<i>Factors to Consider</i>
People/population	Studies involving people on long-haul flights	Definition of long haul Risk status of participants in primary studies Anyone At high risk for DVT (e.g., past history of DVT)
Intervention	Compression stockings	How is “compression stockings” defined? Minimum level of pressure at the ankle? Minimum level of pressure below knee or full leg? As defined by study authors?
Comparator(s)	No compression stockings	Alternative comparators may be: Stockings providing different pressures Aspirin In-flight exercise program.
Outcome	DVT events	How will these be detected in eligible studies? Symptomatic DVTs only or screening for silent DVTs? If screening, which methods are valid? –
Study Design	RCTS only	May want to exclude RCTs at risk of bias No blinded outcome assessment No allocation concealment
Language	Any language	Are translation resources available?
Publication status	Published or unpublished	Will conference abstracts be included?

Note: DVT = deep vein thrombosis. RCT = randomized controlled trial.

How outcomes will be handled in study selection is the subject of much debate; there is a strong argument for study selection not being influenced by whether particular outcomes are reported (Higgins & Green, 2011). Recent research has shown the seriousness of outcome reporting bias, such that study authors sometimes do not report the result for a particular outcome even if they measured it. If reviewers then choose to exclude studies from reviews because an outcome is not reported this compounds the bias (Kirkham et al., 2010). One strategy is to seek out original trial protocols via trial registers such as www.clinicaltrials.gov to determine if the investigators originally intended to measure the outcome of interest and then write to the study authors to request access to the data. AHRQ (2014) offers some useful detailed guidance on defining eligibility criteria. Where resources allow it is important not to limit study inclusion by the language of publication since studies with statistically significant results are more likely to be published in English (and therefore to only include articles in English risks a biased conclusion). There is a similar bias associated with articles that get published at all (Higgins & Green, 2011, Section 10).

Search strategy

Devise a search strategy. Reviewers are advised to work with information scientists (librarians) expert in designing and running maximally sensitive search strategies in the health sciences literature. Typically a search strategy will address all the elements outlined in Table 23.1 (including issues related to outcomes, discussed above). A basic search strategy can be tailored to fit particular databases. The information scientist will also advise which databases should be searched; this will be influenced by the nature of the topic since some topics are better covered by particular databases (CRD, 2009; Higgins & Green, 2011, Section 6).

Selection process

Plan how to select studies for inclusion from your search results and how to document inclusion and exclusion decisions. Aim to minimize bias and mistakes and keep careful records. Ideally, at least two qualified reviewers undertake the selection process based on the protocol. The selection process involves (a) screening titles and abstracts from the searches of various databases to identify potentially eligible studies and (b) obtaining full text copies of potentially eligible studies to screen in detail against the eligibility criteria. Masking reviewers to the identity of original study authors during study selection increases workload without reducing bias (Berlin, 1997). In accord with internationally accepted Preferred Reporting Items for Systematic Reviews and Meta-Analyses recommendations (PRISMA; Liberati et al., 2009), the total number of study reports identified by the electronic search; total selected for full text screening (and the number discarded); total number of studies selected for inclusion in the review after full text screening (and the number discarded with a reason for each) should be recorded and included in the report of your review.

Data extraction

Identify the data elements that will be extracted from included studies, including information relevant to assessment of potential bias. The term “risk of bias” is preferred to “quality” when referring to the strength of the studies individually and as a group because it refers specifically to threats to internal validity. By contrast, “quality” is a term that encompasses other features of the evidence including the precision of any results, which themselves are driven by the amount of data (overall sample size). To ensure accuracy in data extraction, at least two reviewers are essential. A means of extracting and organizing data should be devised (e.g., into a spreadsheet or

a word-processing package or onto paper). An approach to assuring *systematic* assessment of the risk of bias in the primary research studies should be included, so variables related to risk of bias and their definitions should be included in the protocol.

Data analysis plan

Make a data analysis plan including the key measure of treatment effect. When a primary outcome is binary (e.g., hospital-acquired infection or not), treatment effect may be indexed using a risk ratio or odds ratio (Higgins & Green, 2011). When a primary outcome is a continuously measured variable (e.g., body weight), the measure of effect is likely to be a difference in group means.

The analysis plan should be driven by the fundamental review questions, not by the data eventually extracted. Comparisons are driven by the primary objective; secondary questions may address subgroup analysis, for example. The data analysis plan includes decisions about whether to pool data from different studies (meta-analysis) and how to pool them (e.g., by using fixed- or random-effects models). Exploration and assessment of heterogeneity among studies should be included. All these decisions should be justified in the data analysis plan.

Dissemination plan

Complete protocols include a dissemination plan to assure that findings reach the target audience. Because treatment effects may change over time, a plan for updating the review in light of future findings should be included.

Protocol peer review

Protocols should be peer reviewed and made available on websites or in publications. Review protocols produced within the Cochrane Collaboration are formally peer reviewed as part of the process and then published in the Cochrane Library (www.thecochranelibrary.com) whereupon further open comment and feedback is encouraged. Most reviews are improved by seeking feedback from others. Systematic reviewers who are not part of a formal organization that requires peer review may themselves organize protocol reviews amongst peers.

Prospective registration in an international register

Formal, prospective registration of clinical trials in registers such as www.clinicaltrials.gov has long been recognized as an essential step towards eradicating publication bias and ensuring transparency and full disclosure of primary trial results (Dwan, Gamble, Williamson, Kirkham, & Reporting Bias Group, 2013). Prospective registration of systematic reviews is now advocated for similar reasons as well as for efficiency (i.e., reducing unintended duplication of reviewing activity; Stewart, Moher, & Shekelle, 2012). The main international register for systematic reviews is PROSPERO (www.crd.york.ac.uk/prospero/), which incorporates all Cochrane protocols as well as those submitted by reviewers working on non-Cochrane reviews.

Conducting the systematic review

Steps outlined in the published protocol are used to conduct the review. Suggestions for implementation at each step are listed in Table 23.2. Decisions and findings, including search results, are carefully documented. Deviation from the prespecified protocol should be avoided. Any

Table 23.2 Systematic review: implementing protocol steps

Step	Guidance
Searching for eligible studies	Follow the PRISMA flowchart (www.prisma-statement.org/) Keep careful records to ensure a complete PRISMA flowchart
Selecting eligible studies	Follow the protocol Tally reasons for exclusion at each step (following assessment of titles, abstracts, and full-text reads)
Extracting data	Follow the protocol Avoid slavish copying of all participant baseline characteristics and results To avoid reporting bias, avoid extracting data just because they are available (Kirkham et al., 2010) Use the Cochrane Risk of Bias tool to articulate reasoning about judgment bias (Higgins et al., 2011)
Analyzing data	Select software for analyzing extracted data in advance to anticipate how data will need to be handled (RevMan; http://ims.cochrane.org/revman) (Comprehensive Meta Analysis © Biostat Inc or Stata © StataCorp LP; Sterne, Bradburn, & Egger, 2008)
Interpreting results	Consider numeric results in light of risks of bias at study and outcome levels (Higgins et al., 2011) See PRISMA guidelines for Discussion Sections of systematic reviews (Liberati et al., 2009) See the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system for structuring results suitable for the development of guidelines (Guyatt et al., 2011)
Disseminating results	Identify target audiences in advance, for example Academic publication in a high impact journal Press release Social media (Carpenter, Nieva, Albaghal, & Sorra, 2005; Terras, 2012; Wilson, Petticrew, Calnan, & Nazareth, 2010) Identify the best way to reach target audiences

Note: PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

deviations arising from an unforeseen comparison should be documented, clearly labeled as *post hoc*, and justification provided. Decisions that could introduce bias should be reconsidered. For example, a decision to analyze and report a surrogate outcome measure may be considered if the outcome identified in the protocol was used in few studies; however, there is ample evidence that surrogate outcomes can be extremely misleading (Svensson, Menkes, & Lexchin, 2013).

Meta-analysis

Meta-analysis is a statistical procedure that combines the results of separate research studies (Field & Gillett, 2010; also see Cochran, 1954). At least two studies similar in clinical (PICO) and experimental (design) senses are needed to proceed. Meta-analysis involves pooling data across studies to obtain more precise estimates of comparative treatment effects. Evaluation of heterogeneity

(variation) of findings across studies may allow identification of factors in study design and implementation that may have influenced treatment effect estimates in primary studies. Potential moderators of effect size include instrumentation used to measure the primary outcome; time following delivery of intervention at which the outcome was measured; patient demographics; and aspects of intervention components (Coffman, this volume; Sidani & Fleury, this volume). Thus, meta-analysis makes maximum use of available information to better inform strategies in health care. Meta-analysis can be conducted using individual participant data (IPD) from each study or from aggregate data extracted from primary research reports. We emphasize aggregate (study-level) data approaches as these are the most frequently used.

Individual participant data approach

Meta-analysis can be conducted using IPD from each study (Riley, Lambert, & Abo-Zaid, 2010), but doing so involves collaborating with original study author-investigators of all relevant studies to gain access to the patient-level data so it can be a much more time consuming and involved process. Accounting for clustering of IPD within studies is needed to avoid misleading conclusions about effect size in meta-analysis (Abo-Zaid et al., 2013; also see Grittner & Lahmann, this volume). IPD may be combined with aggregate data when IPD are not available from all studies (Riley et al., 2008).

Aggregate data approach

Traditional methods of meta-analysis synthesize aggregate findings extracted from study reports. The nature of the outcome (event or continuously measured health status indicator) needs to be considered before combining the data and estimating the pooled effect size.

Outcomes and effect sizes

Effect size estimates from individual studies are the dependent variable in a meta-analysis. The nature of the primary outcome variable determines which effect size estimate is used; some examples are continuously measured outcomes such as systolic blood pressure, event occurrences such as presence of a decubitus ulcer, or counts such as number of falls. Commonly used effect size measures for continuous and dichotomous outcome data and time to event data are listed in Table 23.3. It is also possible to pool rate and ordinal outcome data but this is less common and would require specialist statistical input.

Estimating the pooled effect size and standard error

The pooled effect size is a weighted overall treatment effect across studies. The weight accorded each study is determined by the amount of information each study contributes to the pooled estimate of the effect. The quality of individual studies and reports varies (e.g., Guo, Sward, Beck, & Staggers, 2014) but study quality is not used to weight the contributions of individual studies to estimates of pooled effect size because precisely how study quality influences study results is not known. Instead, the inverse of the variance of the effect size estimate is used so that primary studies with larger sample sizes have greater weighting in a meta-analysis. Significance levels or confidence intervals are used to gauge uncertainty in pooled estimates of effect size.

Nature of Data		Summary Measure		Examples Data Type
Effect Measure	Expression (for a two arm randomized controlled trial with Experiment group [E] and Control group [C])	Interpretation		
Continuous	Mean difference $Mean\ value\ (E) - Mean\ Value\ (C)$	The difference between the mean values in one group relative to the other. No difference between groups (the null) is zero. Continuous data from different scales can be meta-analyzed if they are on a uniform scale or can be standardized.	- Blood pressure - Body weight	
Dichotomous	Odds ratio $\frac{Odds\ of\ an\ event\ (E)}{Odds\ of\ an\ event\ (C)}$ Where the odds of event in a trial arm = $\frac{Number\ of\ people\ with\ the\ outcome\ in\ the\ group}{Number\ of\ people\ without\ the\ outcome\ in\ the\ group}$	Reports the increased (or decreased) odds of an event in one group compared to another. No difference between groups (the null) is 1. An odds ratio of 2 can be interpreted as the odds of the event being twice as likely in one group compared to the other. The odds ratio has appealing mathematical properties but it is more difficult to interpret than the risk ratio.	- Yes/no - Alive/dead - Cured/not cured - Healed/not healed	
	Risk ratio (relative risk) $\frac{Risk\ of\ event\ in\ (E)}{Risk\ of\ event\ in\ (C)}$ Where the risk of event in each trial group = $\frac{Number\ of\ people\ with\ outcome\ in\ the\ group}{Total\ number\ in\ the\ Group\ (n)}$	Reports the increased (or decreased) risk of an event in one group compared to another. No difference between groups (the null) is 1. Interpretation of risk ratios is straight forward as they are calculated using probabilities. A relative risk of 2 suggests that events in one group are twice as likely as the other.		
Time to Event	Hazard ratio $\frac{Hazard\ of\ the\ event\ in\ (E)\ at\ time\ t}{Hazard\ of\ the\ event\ in\ (C)\ at\ time\ t}$ Where the hazard (at a specific time) is the probability that a participant being observed at time t has the event of interest at that time (given that they have not had event up to point t)	The hazard ratio describes how many times more or less likely an event is in one group compared to the other (at a particular point in time). Interpretation is similar to the risk ratio but the hazard is an instantaneous risk and can change over time (where as the risk ratio is a cumulative risk). However, to simplify interpretation of the hazard ratio it is usually assumed that the hazard ratio is constant over time (proportional hazards assumption).	- Time to healing - Time to remission or recurrence - Time to death	

Heterogeneity and effect moderators

Heterogeneity refers to the nature and amount of variation in effect size estimates from individual studies included in the meta-analysis. Variation may arise due to differences in study participants (clinical heterogeneity; Gagnier, Moher, Boon, Beyene, & Bombardier, 2012), study design and methods (methodological heterogeneity), and/or statistical heterogeneity (sampling variation) (Song, Sheldon, Sutton, Abrams, & Jones, 2001). Heterogeneity should initially be informally assessed to determine whether effect sizes appear similar in direction and magnitude and then formally assessed using statistical methods (e.g., Raudenbush & Bryk, 2002, Chapter 7). The statistical approach involves two steps. First, analysis is conducted to determine if observed heterogeneity in effect size is significantly greater than that expected due to sampling error. If the null hypothesis that all observed effect size variation is due to sampling variability is rejected, then further analyses to address the extent and causes of heterogeneity in study effect size are conducted.

As an example, we conducted a meta-analysis of five RCTs comparing use of a standard hospital mattress with a foam alternative based for prevention of pressure ulcer on a total of 2,016 participants from five studies (McInnes, Jammali-Blasi, Bell-Syer, Dumville, & Cullum, 2011, Analysis 2.1, Analysis 2.2). We constructed a forest plot to display results and computed I^2 and χ^2 statistics to index heterogeneity in the results from the five studies.

Forest plots are graphical representations of the findings from the complete set of individual studies included in a meta-analysis. As shown in Figure 23.2, the point estimate of effect size and confidence interval are shown for each study and for the pooled results. To test the extent of statistical heterogeneity, two common measures are used. The first, the χ^2 test for heterogeneity, assesses whether the variation in effect sizes between studies is what would be expected by chance alone. This test is notoriously underpowered in common meta-analysis situations where the number of studies is small, so meta-analysts generally set a lower threshold in their protocol of $p < .10$ for significance (Higgins & Green, 2011). The second approach focuses on the impact of heterogeneity on the meta-analysis by quantifying the inconsistency across studies – that is, the percentage of variation due to heterogeneity rather than chance. This is called I^2 (Higgins et al., 2003). I^2 provides an easy to understand and intuitive measure of heterogeneity of results.

In our meta-analysis (McInnes et al., 2011), risk ratios (RRs) were available for four of the five studies. All RRs were less than 1.0 (indicated fewer pressure ulcers on the alternative foam mattress relative to the standard hospital mattress), of which three were statistically significant. However, tests indicated important heterogeneity ($\chi^2 = 13.24$, $p = .004$; $I^2 = 77\%$; Figure 23.2; for background, see Higgins, Thompson, Deeks, & Altman, 2003). There is much debate about whether it is advisable to pool studies at all in the face of such high levels of heterogeneity. The full review details how the authors (of whom we are two!) investigated possible reasons for the heterogeneity. First, we removed the non-UK study (because the mattresses may be substantially different in different parts of the world); however, I^2 increased to 84% when the non-UK study was removed. We then considered clinical explanations for the high heterogeneity and reasoned that one study counted Grade 1 pressure ulcers (intact skin with local nonblanchable redness) as an event whereas others did not. Removing this study reduced I^2 to 39%, which was nonsignificant. Such exploratory approaches are appropriate within a systematic review because they help to understand what is driving any apparent treatment effect (and whether it might be “real”). However, it is essential that reports include all steps, rationale for them, and results.

Another important decision influenced by heterogeneity is whether to pool using a fixed effect or random effects model. A fixed effect model assumes that all the studies share one underlying, true treatment effect and differences are due to sampling (chance); the fixed effect model ignores heterogeneity (Riley, Higgins, & Deeks, 2011). When obvious heterogeneity exists, a random effects

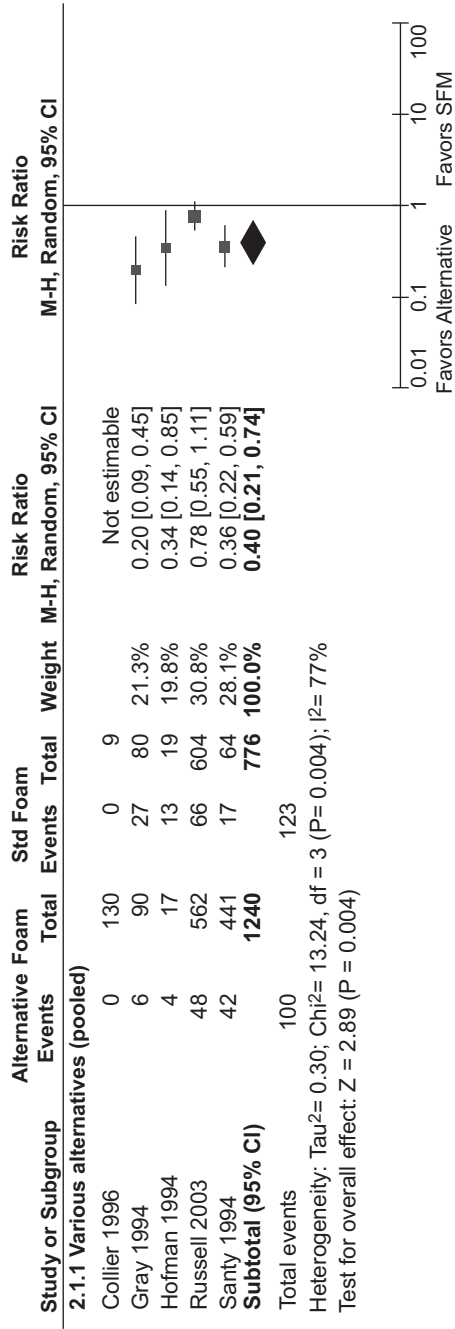


Figure 23.2 Forest plot depicting results from meta-analysis of five randomized controlled trials comparing standard foam mattresses with alternative foam mattresses for the outcome of incident pressure ulcer events. Risk ratios obtained in four of the studies are plotted as shown using boxes; extensions from the boxes represent 95% confidence intervals. The size of the boxes represent the weight assigned to each study (see also the weight column). The center of the diamond shows the pooled estimate of the effect size, and the width of the diamond indicates the 95% confidence interval for the pooled estimate. From McInnes, Jammali-Blasi, Bell-Syer, Dumville, and Cullum (2011).

model that allows for the true effect to vary across studies is used. The random effects model views the studies in a meta-analysis as a random sample of studies with a distribution of results usually assumed to be normal. The random effects model then treats each study as estimating a different treatment effect from this distribution. Individual study weights are less influenced by sample size. Whilst a random effects model can account for heterogeneity, it does not by itself adjust for or explain the source of the heterogeneity. When interpreting results, it is important to bear in mind the implications of using a fixed or random effects model for meta-analysis, but few authors actually do this (Riley et al., 2011). Remember that a fixed effect model estimates a common underlying effect, whereas a random effects model emphasizes identification of a distribution of effects.

Subgroup analysis can also be used to investigate and potentially explain heterogeneity in meta-analysis (Borenstein & Higgins, 2013). Modes of intervention delivery can vary (for example, cognitive behavioral therapy for pain management is delivered by telephone in some studies and face-to-face in others), and creating and analyzing subgroups of trials based on delivery method might identify effect modification; i.e., variation in treatment effect sizes based on mode of intervention delivery. Subgroup analysis may also be undertaken to identify effect modification (sometimes called interactions) between clinically relevant characteristics (such as gender or age) and the treatment effect, where suitable trial data are available. Any subgroup analysis should be undertaken with caution and transparency because they constitute an observational study, with groups not assigned at random. A priori specification of subgroup analysis in a protocol is ideal, but differences may not be apparent until the meta-analysis has been conducted. Subgroup analysis involves splitting the data based on known group membership. The extent to which it is possible to conduct subgroup analysis with aggregate data is severely restricted and depends on the characteristics of the available trials and/or the availability of subgroup findings in study reports. Type I error is inflated when many subgroup analyses are conducted. Readers are strongly urged to read and consult widely before embarking on anything except small-scale, exploratory subgroup analysis (Deeks, Higgins, & Altman, 2011; Fu et al., 2008).

For a highly readable exploration of decision making in meta-analysis and how heterogeneity can and should influence decision making, readers are directed to Ioannidis, Patsopoulos, and Rothstein (2008). What most authors agree on is that heterogeneity should be explored. When this is not possible because the meta-analysis is based on only a few studies, extra care should be taken to emphasize the relative lack of information on which interpretation and conclusions are based. Any differences among study characteristics or samples that may have led to different results should be identified using an informed and commonsense approach. PICO factors should be systematically assessed.

Meta-regression

Effects of study characteristics, such as design features, on estimates of intervention effectiveness may also be investigated using meta-regression (Deeks, Higgins, & Altman, 2011; Thompson & Higgins, 2002). Meta-regression is like simple regression except it uses aggregate level data, with the treatment effect as the outcome and one or more study characteristics that might impact on the treatment effect (potential effect modifiers) as explanatory covariates. Meta-regression involves investigating the effect of a study characteristic (this may be a study design feature such as whether blinded outcome assessment was used or an aspect of PICO) on intervention effects, however, it should not be considered when there are fewer than 10 studies in a meta-analysis (Higgins & Green, 2011, Section 9.6.4). For example, Walsh et al. (2010) used meta-regression to estimate whether extent of tooth decay at baseline and supervision of tooth brushing modified the effects of fluoride toothpaste use on preventing tooth decay in children and adolescents (Walsh et al., 2010). The extent of decay at baseline was incorporated

as an individual-level covariate and supervision of tooth brushing (a protocol variation) as a possible study-level effect modifier.

Software

Simple and complex meta-analysis and meta-regression can be conducted in standard statistical software packages such as SPSS (IBM Corp. Armonk, NY: IBM Corp), STATA (StataCorp, College Station, TX: StataCorp LP), and R (with metafor add-on). Software designed to support the conduct of systematic reviews and meta-analysis is also available such as the Cochrane software RevMan (see <http://ims.cochrane.org/revman>) and Comprehensive Meta-Analysis (Biostat Inc., Englewood, NJ).

Advanced topics in evidence synthesis

Networks of evidence

“Pair-wise” meta-analysis involves synthesis of data from multiple trials evaluating the same intervention and comparator (an intervention/comparator pair). Whilst this process provides important pooled effect estimates for head-to-head comparisons, decision makers often want to answer the broader question: of all the treatment options available which is the best?

For example, compression therapy delivering 40 mmHg (of pressure) at the ankle is a widely used treatment for venous leg ulcers. This high compression improves venous blood flow and reduces time to healing of these chronic wounds. Four commonly used high compression systems (four-layer bandage, short stretch bandage, two-layer hosiery, and two-layer bandage) are available and decision makers need to make a choice about which to use. Four pair-wise comparisons have been evaluated for these treatments in a total of 12 RCTs (Ashby et al., 2014; Figure 23.3).

Decision makers who are choosing only between two treatments can, potentially, rely on relative effect estimates from standard meta-analysis to guide them. Yet, ideally, we want to know which of the four treatment options is likely to be the most effective for treating venous leg ulcers. Making this judgment based on multiple pair-wise data analyses alone is challenging: it requires qualitative assessment of comparisons where no pair-wise RCT evidence may exist as well as a qualitative assessment of the overall network in terms of relative treatment effects and precision of estimates. It is preferable to undertake this assessment quantitatively using meta-analysis extended to maximally synthesize all available RCT evidence. This is called network meta-analysis (sometimes also known as a mixed treatment comparison or mixed treatment meta-analysis).

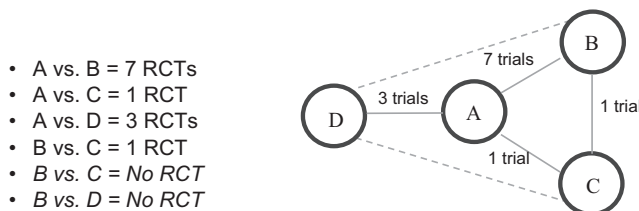


Figure 23.3 Network of evidence of randomized controlled trials (RCTs) for four high compression treatments for venous leg ulcers. Dashed lines show treatments that have not been directly compared in an RCT (Ashby et al., in press). The network contains a data loop that provides direct and indirect evidence for the comparison treatment B vs. C. When a network meta-analysis has transitivity, the direct and indirect estimates for B vs. C are consistent.

Network meta-analysis

Network meta-analysis allows estimation of relative treatment effects for both direct (pair-wise) and indirect comparisons. Estimates for indirect comparisons are calculated using direct pair-wise data that link two treatments indirectly via a common comparator. In our example (Figure 23.3), the direct data from A vs. B and A vs. D can be used to estimate the relative effectiveness of the indirect comparison B vs. D. Similarly direct data for A vs. C and A vs. D allows indirect comparison of C vs. D.

Network meta-analysis has the advantage of using all available and relevant RCT evidence to calculate effect estimates that maximally draw on direct and indirect data whilst, essentially, preserving randomization. Such network meta-analyses can be complex in structure (Barth et al., 2013; Elliott & Meyer, 2007) as well as analytically complex to run; however, they are increasingly common – not least because when run in a Bayesian framework the final output can be a summary of treatment ranked by the percentage of each being the ‘best’ treatment in terms of the outcome being evaluated (Salanti, Ades, & Ioannidis, 2011).

Assumptions in network meta-analysis

Assumptions made in standard meta-analysis must hold across the comparisons in a network meta-analysis. In addition, transitivity of treatment effects must hold (Salanti, 2012): that is, we can make valid estimates of an indirect comparison via a common comparator in a network analysis. Simply stated, this means that if in pair-wise comparisons treatment A was more effective than treatment B and treatment B was more effective than treatment C, then in a three-arm trial comparing treatments A, B, and C, treatment A would be more effective than treatment C. One approach is to assume that the ‘missing’ treatment in each comparison is missing at random (Dias, Welton, Caldwell, & Ades, 2010; Salanti, 2012). Within networks, the assumption of transitivity can be difficult to assess; however, if the assumption holds then where a network contains a data loop (Figure 23.3) and the direct and indirect estimates should be consistent (Cipriani, Higgins, Geddes, & Salanti, 2013).

Quality assessment in network meta-analysis

Network meta-analysis is a developing method. Approaches to quality assessment of evidence produced from network meta-analysis are under development. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach (Guyatt, Oxman, Axl et al., 2011; Guyatt, Oxman, Schünemann, Tugwell, & Knottnerus, 2011) outlines how quality of, or confidence in, estimates from pair-wise comparisons can be rated based on data from several domains and similar processes are now being modified for application to network meta-analysis (e.g., Dumville, Soares, O’Meara, & Cullum, 2012).

Decision analytic modeling

CER and meta-analysis typically focuses on assessment of relative treatment effects based on data obtained from RCTs. Costs and benefits of treatments are increasingly assessed to determine whether healthcare resources are being used to return maximum health benefit (Griffin, Claxton, & Sculpher, 2008). A wide range of evidence requires synthesis when costs and benefits of treatments are assessed (Uchida-Nakakoji & Stone, this volume); decision analytical modeling is used to facilitate this evaluation.

Decision analytic models include key states for a condition or disease in decision tree or Markov models that comprehensively capture key costs and benefits as well as differences between treatment outcomes but without being too complex to implement. A decision analytic model acts as a framework in which data from many sources (e.g., RCTs, meta-analysis or network meta-analysis for estimating relative treatment effects as well as natural history, cost data and quality of life assessments) can be synthesized. Such models can be complex, assumptions need to be made, and critical appraisal is required (Kim & Thompson, 2010; Weinstein et al., 2003).

Conclusion

High-quality comparative effectiveness research is the basis for policy and practice recommendations. Well-stated problems use the patient-intervention-comparator-outcome (PICO) approach. Systematic review and meta-analysis are two steps involved in comprehensive and rigorous synthesis of all the relevant research that addresses a well-stated problem. Networks of evidence, incorporation of quality of studies in systematic review and meta-analysis, and decision analytic modeling are emerging extensions to well-established methods.

References

- Abo-Zaid, G., Guo, B., Deeks, J. J., Debray, T.P.A., Steyerberg, E. W., Moons, G. M., & Riley, R. D. (2013). Individual participant data meta-analyses should not ignore clustering. *Journal of Clinical Epidemiology*, 66, 865–873.
- Agency for Healthcare Research and Quality. (2014, January). Methods guide for effectiveness and comparative effectiveness reviews. AHRQ Publication No. 10(14)-EHC063-EF Washington, DC: Author. Retrieved from <http://effectivehealthcare.ahrq.gov/ehc/products/60/318/CER-Methods-Guide-140109.pdf>
- Agency for Healthcare Research and Quality. (n.d.). Glossary of terms. Retrieved from <http://effectivehealthcare.ahrq.gov/index.cfm/glossary-of-terms/?filterletter=c>
- Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA*, 268, 240–248.
- Ashby, R. L., Gabe, R., Ali, S., Saramago, P., Chuang, L.-H., Adderley, U., . . . Torgerson, D. J. (2014). VenUS IV (Venous leg Ulcer Study IV): Compression hosiery versus compression bandaging in the treatment of venous leg ulcers: A randomised controlled trial, mixed treatment comparison and decision analytic model. *Health Technology Assessment*, 18(57), 1–293, vi–vi.
- Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H., . . . Cuijpers, P. (2013). Comparative efficacy of seven psychotherapeutic interventions for patients with depression: A network meta-analysis. *PLoS Medicine*, 10(5), e1001454.
- Berlin, J. A. (1997). Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. *Lancet*, 350(9072), 185–186.
- Borenstein, M., & Higgins, J.P.T. (2013). Meta-analysis and subgroups. *Prevention Science*, 14, 134–143.
- Carpenter, D., Nieva, V., Albaghal, T., & Sorra, J. (2005). Development of a planning tool to guide research dissemination. In K. Henriksen, J. B. Battles, E. S. Marks, & D. I. Lewin (Eds.), *Advances in patient safety: From research to implementation (Volume 4: Programs, tools, and products)*. Rockville, MD: Agency for Healthcare Research and Quality (US). Retrieved from www.ncbi.nlm.nih.gov/books/NBK20603/
- Centre for Reviews and Dissemination, University of York. (2009). Systematic reviews. CRD's guidance for undertaking reviews in health care. Retrieved from www.york.ac.uk/inst/crd/SysRev/!SSL!/WebHelp/SysRev3.htm
- Chalmers, I., Hetherington, J., Newdick, M., Mutch, L., Grant, A., Enkin, M., . . . Dickersen, K. (1986). The Oxford Database of Perinatal Trials: Developing a register of published reports of controlled trials. *Controlled Clinical Trials*, 7, 306–324.
- Cipriani, A., Higgins, J.P.T., Geddes, J.R., & Salanti, G. (2013). Conceptual and technical challenges in network meta-analysis. *Annals of Internal Medicine*, 159, 130–137.

- Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101–129.
- Cochrane, A. L. (1979). 1931–1971: A critical review, with particular reference to the medical profession. In G. Teeling-Smith & N. Wells (Eds.), *Medicines for the year 2000* (pp. 1–11). London, UK: Office of Health Economics.
- Cochrane Collaboration. (2013). History. Retrieved from www.cochrane.org/about-us/history
- Deeks, J. J., Higgins, J.P.T., & Altman, D.G. (2011, March). 9.6.4. Meta-regression. In J.P.T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions, version 5.1.0*. Retrieved from <http://handbook.cochrane.org/>
- Dias, S., Welton, N. J., Caldwell, D.M., & Ades, A.E. (2010). Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine*, 29, 932–944.
- Dickersen, K. (1986). The Oxford database of perinatal trials: Developing a register of published reports of controlled trials. *Controlled Clinical Trials*, 7, 306–324.
- Droogan, J., & Cullum, N. (1998). Systematic reviews in nursing. *International Journal of Nursing Studies*, 35(1–2), 13–22.
- Dumville, J.C., Soares, M.O., O'Meara, S., & Cullum, N. (2012). Systematic review and mixed treatment comparison: Dressings to heal diabetic foot ulcers. *Diabetologia*, 55, 1902–1910.
- Dwan, K., Gamble, C., Williamson, P.R., Kirkham, J. J., & Reporting Bias Group. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias – an updated review. *PloS One*, 8(7), e66844.
- Elliott, W. J., & Meyer, P. M. (2007). Incident diabetes in clinical trials of antihypertensive drugs: A network meta-analysis. *Lancet*, 369(9557), 201–207.
- Field, A.P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 63, 665–694.
- Fineout-Overholt, E., & Johnston, L. (2005). Teaching EBP: Asking searchable, answerable clinical questions. *Worldviews on Evidence-based Nursing*, 2, 157–160.
- Fu, R., Gartlehner, G., Grant, M., Shamliyan, T., Sedrakyan, A., Wilt, T. J., . . . Trikalinos, T. A. (2008). Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. In *Methods guide for effectiveness and comparative effectiveness reviews*. Rockville, MD: Agency for Healthcare Research and Quality (US). Retrieved from www.ncbi.nlm.nih.gov/books/NBK49407/
- Gagnier, J. J., Moher, D., Boon, H., Beyene, J., & Bombardier, C. (2012). Investigating clinical heterogeneity in systematic reviews: A methodological review of guidance in the literature. *BMC Medical Research Methodology*, 12(111). doi:10.1186/1471-2288-12-111
- Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, 5, 351–379.
- Griffin, S., Claxton, K., & Sculpher, M. (2008). Decision analysis for resource allocation in health care. *Journal of Health Services Research and Policy*, 13(Suppl. 3), 23–30.
- Guo, J. W., Sward, K. A., Beck, S. L., & Staggers, N. (2014). Quality of reporting randomized controlled trials in cancer nursing research. *Nursing Research*, 63, 26–35.
- Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., . . . Schünemann, H. J. (2011). GRADE guidelines: 1. Introduction – GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, 64, 383–394.
- Guyatt, G. H., Oxman, A. D., Schünemann, H. J., Tugwell, P., & Knottnerus, A. (2011). GRADE guidelines: A new series of articles in the *Journal of Clinical Epidemiology*. *Journal of Clinical Epidemiology*, 64, 380–382.
- Higgins, J.P.T., Altman, D.G., Gotzsche, P.C., Jüni, P., Moher, D., Oxman, A.D., . . . Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ (Clinical Research Ed.)*, 343, d5928.
- Higgins, J.P.T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions*. Version 5.1.0. The Cochrane Collaboration. Available from www.cochrane-handbook.org
- Higgins, J.P.T., Thompson, S.G., Deeks, J. J., & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414), 557–560.
- Hill, G. B. (2000). Archie Cochrane and his legacy. An internal challenge to physicians' autonomy? *Journal of Clinical Epidemiology*, 53, 1189–1192.
- Institute of Medicine. (2011). Standards for developing trustworthy clinical practice guidelines. Retrieved from www.iom.edu/Reports/2011/Clinical-Practice-Guidelines-We-Can-Trust/Standards.aspx
- Ioannidis, J.P.A., Patsopoulos, N.A., & Rothstein, H.R. (2008). Reasons or excuses for avoiding meta-analysis in forest plots. *BMJ*, 336(7658), 1413–1415.

- Kim, L. G., & Thompson, S. G. (2010). Uncertainty and validation of health economic decision models. *Health Economics*, 19, 43–55.
- Kirkham, J. J., Dwan, K. M., Altman, D. G., Gamble, C., Dodd, S., Smyth, R., & Williamson, P. R. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*, 340, c365.
- Liberati, A., Altman, D., Tetzlaff, J., Mulrow, C., Gotzsche, P., Ioannidis, J., . . . Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine*, 6(7), e1000100.
- McInnes, E., Jammali-Blasi, A., Bell-Syer, S. E., Dumville, J. C., & Cullum, N. (2011). Support surfaces for pressure ulcer prevention. *Cochrane Database of Systematic Reviews (Online)*, 4. Retrieved from www.scopus.com/inward/record.url?eid=2-s2.0-79955669600&partnerID=40&md5=3b87160eecd70feb96225ceb59c25f6
- Methods Guide for Effectiveness and Comparative Effectiveness Reviews. (2014). Rockville, MD: Agency for Healthcare Research and Quality. Retrieved from <http://effectivehealthcare.ahrq.gov/ehc/products/60/318/CER-Methods-Guide-140109.pdf>
- Moseley, A. M., Elkins, M. R., Herbert, R. D., Maher, C. G., & Sherrington, C. (2009). Cochrane reviews used more rigorous methods than non-Cochrane reviews: Survey of systematic reviews in physiotherapy. *Journal of Clinical Epidemiology*, 62, 1021–1030.
- Mulrow, C. D. (1987). The medical review article: State of the science. *Annals of Internal Medicine*, 106, 485–488.
- O’Flynn, A. I. (1982). Meta-analysis. *Nursing Research*, 31, 314–316.
- Olsen, O., Middleton, P., Ezzo, J., Gotzsche, P. C., Hadhazy, V., Herxheimer, A., . . . McIntosh, H. (2001). Quality of Cochrane reviews: Assessment of sample from 1998. *BMJ*, 323(7317), 829–832.
- Raudenbush, S. W., & Bryk, A. S. (2002). Applications in meta-analysis and other cases where level-1 variances are known. In *Hierarchical linear models* (2nd ed., pp. 205–227). Thousand Oaks, CA: Sage.
- Riley, R. D., Higgins, J. P. T., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ*, 342, d549.
- Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ*, 340, c221. doi:10.1136/bmj.c221
- Riley, R. D., Lambert, P. C., Staessen, J. A., Wang, J., Gueyffier, F., Thijs, L., & Bouillon-Bu, F. (2008). Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine*, 27, 1870–1893.
- Sabharwal, S., Patel, V., Nijjer, S. S., Kirresh, A., Darzi, A., Chambers, J. C., . . . Athanasiou, T. (2013). Guidelines in cardiac clinical practice: Evaluation of their methodological quality using the AGREE II instrument. *Journal of the Royal Society of Medicine*, 106, 315–322.
- Salanti, G. (2012). Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: Many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods*, 3, 80–97.
- Salanti, G., Ades, A. E., & Ioannidis, J. P. A. (2011). Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: An overview and tutorial. *Journal of Clinical Epidemiology*, 64, 163–171.
- Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32, 752–760.
- Song, F., Eastwood, A. J., Gilbody, S., Duley, L., & Sutton, A. J. (2000). Publication and related bias. *Health Technology Assessment*, 4(10). Retrieved from www.hta.ac.uk/fullmono/mon410.pdf
- Song, F., Sheldon, T. A., Sutton, A. J., Abrams, K. R., & Jones, D. R. (2001). Methods for exploring heterogeneity in meta-analysis. *Evaluation & the Health Professions*, 24, 126–151.
- Sterne, J. A. C., Bradburn, M. J., & Egger, M. (2008). Meta-analysis in Stata™. In M. Egger, G. D. Smith, & D. G. Altman (Eds.), *Systematic reviews in health care* (2nd ed., pp. 347–369). London, UK: BMJ. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/9780470693926.ch18/summary>
- Stewart, L., Moher, D., & Shekelle, P. (2012). Why prospective registration of systematic reviews makes sense [Editorial]. *Systematic Reviews*, 1, 7.
- Svensson S., Menkes, D. B., & Lexchin, J. (2013). Surrogate outcomes in clinical trials: A cautionary tale. *JAMA Internal Medicine*, 173, 611–612.
- Terras, M. (2012). The impact of social media on the dissemination of research: Results of an experiment. *Journal of Digital Humanities*, 1(3). Retrieved from <http://journalofdigitalhumanities.org/1-3/the-impact-of-social-media-on-the-dissemination-of-research-by-melissa-terras/>

- Thompson, S.G., & Higgins, J.P.T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, 21, 1559–1573.
- Walsh, T., Worthington, H. V., Glenny, A.-M., Appelbe, P., Marinho, V.C., & Shi, X. (2010). Fluoride tooth-pastes of different concentrations for preventing dental caries in children and adolescents. In *Cochrane Database of Systematic Reviews*. Chichester: Wiley. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD007868.pub2/abstract>
- Weinstein, M.C., O'Brien, B., Hornberger, J., Jackson, J., Johannesson, M., McCabe, C., . . . ISPOR Task Force on Good Research Practices – Modeling Studies. (2003). Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ISPOR Task Force on Good Research Practices – Modeling Studies. *Value Health*, 6, 9–17.
- Wilson, P. M., Petticrew, M., Calnan, M. W., & Nazareth, I. (2010). Disseminating research findings: What should researchers do? A systematic scoping review of conceptual frameworks. *Implementation Science*, 5, 91.

PATIENT SAFETY RESEARCH

Methodological challenges

Sean P. Clarke and Maria Schubert

The current international health care safety movement is often considered to have started with publication of the Institute of Medicine report *To Err Is Human* (Institute of Medicine [IOM], 2000). This call to reduce errors and associated morbidity and mortality, bolstered by shocking statistics, galvanized clinicians, researchers, and health care leaders. Despite slim mention of nursing in the report, nurse researchers heeded the call, especially since *To Err Is Human* appeared on the heels of widespread restructuring in Western health care systems that often led to reduced nurse staffing.

“Safety” is a flashpoint word – it signals urgency and calls to mind dramatic incidents. Safety research in nursing is generally directed towards reducing untoward incidents for patients/users of health services that have potential implications ranging from short-term or long-term complications and suffering to death. In this chapter, safety research refers primarily to generation of empirical data to improve consistency and accuracy of nursing care delivery and avoid adverse events (AEs). Research questions and designs are described, a framework is outlined, challenges are identified, and future directions for safety research in nursing are outlined.

Typical questions and designs in safety research in nursing

Research questions addressed in safety research in nursing tend to fall into a few categories: (a) descriptive – related to definitions and estimation of prevalence and severity of problems; (b) correlational – focused on associations between safety issues and client, provider, and system factors; and (c) evaluative – aimed at developing and gauging impacts of interventions to improve safety. The same quantitative and qualitative approaches used in other nursing research are also used in safety studies. Observational studies (Harrison, this volume) can be used to study various aspects of patient care (for instance, evaluate interruptions during medication administration; Merwin & Thornlow, 2006). Quasi-experiments are used (Eckardt & Rindskopf, this volume); for instance, pretest, posttest comparative design to test the impact of implementing a tool, protocol, or a system intervention such as executive walking rounds on patient safety or quality of care.

Quantitative data are collected through structured interviews or surveys (the latter especially to measure safety climate or system factors; e.g., quality of the nurse work environment). Secondary data (collected for another purpose) can also be used. For instance, administrative data or patient discharge abstracts and structured reviews of records and charts can be used to understand why patients presented for care, what care was received, what the health status of patients was at

different points in time, and what negative events occurred while patients were in care. Secondary data can also be used to identify mortality or AE rates or to gauge costs of care. Rating grids can be used to compare practices against a standard.

Qualitative approaches include individual and focus group interviews to better understand perceptions of and circumstances surrounding issues related to errors or AEs, and ethnographic studies to evaluate introduction of new safety systems. Another option is to directly observe care or act as a participant–observer in a setting.

Mixed-methods safety research is increasingly common. In mixed-methods studies, a variety of strategies that draw upon the strengths of the quantitative and qualitative traditions to address complex research questions and multifaceted phenomena can be undertaken (Creswell & Plano Clark, 2011; Doyle, this volume; Harris et al., 2005). For instance, Robert et al. (2011) combined qualitative research methods, including an ethnographic case study with interviews and direct nonparticipant observation of organizational processes, with quantitative research methods (AE and patient complaint data to explore relationships between organizational and cultural characteristics of hospitals and their impacts on the quality of health care.

Key concepts in safety research

Errors, mistakes, violations, and AEs are central concepts to safety research. The notions of antecedents and root causes are used to understand linkages among these ideas, and are essential to designing and evaluating interventions.

Event types

Patient safety is conventionally defined as freedom from accidental injuries due to medical care or medical errors (Freeth et al., 2012; IOM, 2000). Practices or interventions to improve patient safety have often targeted reduction of preventable AEs in hospitalized patients, though they are increasingly aimed at patients and services across the continuum of care (e.g., outpatient and community settings, long-term care facilities). Errors, AEs, and critical incidents are occurrences connected with safety; each has a distinct definition in research but in many clinicians' minds and the popular imagination, they are part of a larger, somewhat amorphous group of negative events in health care settings.

Table 24.1 lists key event types in safety research. Errors, mistakes, and violations are care provider actions, whereas AEs are outcomes of care as experienced by patients. Errors and mistakes include lapses, slips, trips, fumbles, rule-based mistakes, and knowledge-based mistakes. Mistakes may be related to timing, operations, therapeutic processes, checking and verifications during procedures, information use, and selection of courses of actions. Violations are deviations from standard operating practices, procedures, standards, or rules. The word “violation” implies deviant behavior that merits policing and, perhaps, even punishment. If an action is indeed a breach of a relevant established rule, then violation is likely the correct term. However, it should be kept in mind that at times rule bending can be necessary for patient safety under circumstances more complicated than those accounted for by authors of rules (Kreckler, Catchpole, Bottomley, Handa, & McCulloch, 2008). Also, rules change over time as cultures of practice and science evolve.

AEs (therapeutic misadventures) are harmful events, accidents, or injuries due to treatment rather than underlying diseases; AEs range from insignificant to serious or life-threatening (Table 24.1). AEs are directly associated with care or services provided – but may or may not be the result of errors or mistakes. AEs include side effects of diagnostic procedures or medical treatments, therapeutic mishaps, iatrogenic injuries, or other events that may or may not be preventable. AE impacts

Table 24.1 Errors, mistakes, adverse events, and violations

<i>Problem</i>	<i>Definition</i>	<i>Subtype</i>	<i>Characteristics</i>
Error	Unintentional Incident/event ^a Failure of executions	• Lapse • Slip	• Failure of execution arising from memory loss • Attentional or perceptual failures of execution arising from distractions or preoccupations
Mistake	Actions arising from faulty/incomplete knowledge ^b	• Rule-based mistake • Knowledge-based mistake • Timing^c • Operational^a • Therapeutic process^a • Checking^a • Retrieval^a • Communication • Selection	• Incorrect application or selection of a solution; a correct action is selected for the wrong problem • Planning or problem-solving failure involving inadequate plans, often occurring in new situations without rules or preprogrammed solutions • Correct or incorrect action is carried out too soon or too late • Correct action is carried out too slowly or quickly • Faulty history taking, mistaken diagnosis, prescribing error • Verifications are omitted, incomplete, or carried out on wrong objects • Correct information not obtained, or wrong/incomplete information obtained • Information exchanged is wrong, incomplete, or misunderstood • Inappropriate intervention choice is made
Adverse event	Harmful event, accident or injury due to treatment rather than underlying disease	• Significant AE • Serious AE • Life-threatening AE^c	• AE not resulting in permanent change in functioning • AE resulting in permanent change in functioning • AE resulting in high risk of death or death
Violation	Deviation from standard operating practices, procedure standards, or rules	• Optimizing • Necessary/situational • Routine	• Actions taken to further personal purposes rather than goals related to the task at hand • Deviations taken to deal with contingencies of particular situations • Cutting corners

Note: AE = Adverse event.

^aVincent (2010). ^bKreckler, Catchpole, Bottomley, Handa, and McCulloch (2008). ^cKale, Keohane, Maviglia, Gandhi, and Poon (2012).

vary somewhat by type of event. For example, certain medication administration errors (wrong dose, wrong medication, wrong route of administration, improper preparation), administering a contraindicated medication (e.g., one to which a patient has a documented allergy), or delivery of an incorrect treatment following a wrong diagnosis tend to be associated with serious or life-threatening

AEs (Hoffmann & Rohe, 2010; Kale, Keohane, Maviglia, Gandhi, & Poon 2012; IOM 2004). Most AEs (62.5% to 83%) have been judged potentially preventable through specific interventions (Curry, Nembhard, & Bradley, 2009; von Laue, Schwappach, & Koeck, 2003). Attributing causes for errors and omissions in care (or deeming an AE unlikely to have been preventable) and identifying the impact of the AE on a patient's health may be a complicated theoretical (or even philosophical) questions. As distinctions involving some subjectivity, they can pose considerable methodological challenges.

Antecedents and root causes

The occurrence of AEs prompts a search for explanations in terms of root causes (mechanisms) that have led to possible errors, mistakes, and violations. Causes may be latent or active. Latent failures are underlying conditions associated with individuals (e.g., inexperience, inadequate training, emotional or physical status) or work environments (e.g., heavy workload, poor communication or collaboration, outdated equipment; Reason, 1990); they are preconditions for unsafe acts arising from technical and organizational actions and decisions often made long before an accident takes place (Lawton et al., 2012; von Laue et al., 2003). Latent failures may affect performance of clinicians, creating active failures that lead to slips, lapses, or mistakes and result in AEs, failing defense mechanisms to stop the process. Active failures are shortcomings of performance or behavior, at the “sharp end” of patient care (so called because errors sometimes relate to the handling of instruments; however, the term “sharp end” refers to any direct contact with patients).

Post-hoc research on errors and AEs is challenging because many errors are relatively uncommon, it is often impossible to objectively assess conditions in place at the time of an incident, complete information about all potentially relevant circumstances is never available, and data available to reviewers after an event or error may not have been on hand at the time clinical decisions were made. When those reviewing an incident are aware of its outcome (lasting harm to the patient, changes in the patient's treatment trajectory), ratings of error severity and attribution of causes in terms of behaviors or acts may change (De Meester, Van Bogaert, Clarke, & Bossaert, 2013; Henriksen & Kaplan, 2003). Cross-sectional correlational studies examining rates of AEs or errors in relation to environmental conditions provide the main support for the notion of latent errors and rely on merging information from patient and management data files to recreate conditions of events.

Factors associated with adverse events and errors

Human and system factors are associated with errors and AEs. Examples include connections between workload, patient-to-nurse ratio, and errors in general (Meurier, 2000; Weissman et al., 2007), and more specifically between nursing workload (RN skill mix, total nursing care hours) and medication administration errors, falls without and with injuries, and inpatient mortality (Aiken, Clarke, Sloane, Sochalski, & Silber, 2002; Balas, Scott, & Rogers, 2004; Cho & Yun, 2009; Diya, Van den Heede, Sermeus, & Lesaffre, 2012; Kane, Shamliyan, Mueller, Duval, & Wilt, 2007; Patrician et al., 2011; Twigg, Duffield, Bremner, Rapley, & Finn, 2011). Communication (adequate/inadequate), collaboration and/or teamwork among health care providers, support (or lack of support) from nursing leaders, and empowerment have been linked with patient safety and/or etiology and prevention of errors and near errors (Balas et al., 2004; Meurier, 2000; Richardson & Storr, 2010). Some researchers have presented evidence for an association between safety culture or climate and AEs and error rates (Mardon, Khanna, Sorra, Dyer, & Famolaro, 2010; Singer, Lin, Falwell, Gaba, & Baker, 2009; Vogus & Sutcliffe, 2007). However, use of different measures for similar concepts, relatively small effect sizes, and negative findings limit confidence in these conclusions.

Team context is vital in health care settings. Consensus is that effective functioning of interdisciplinary health care teams that come together to provide services is essential and miscommunication underlies many safety issues. The degree to which members of health care teams are able to understand each other's roles and collaborate to reach common goals varies across contexts (e.g., Wagner, Smits, Sorra, & Huang, 2013). Detailed understanding of team functioning needed to guide development of educational and other interventions is still lacking. Multilevel models (Grittner & Lahmann, this volume) and models for intensive longitudinal data (DeBoeck & Boker, this volume) may be useful for investigating these issues.

Interventions

Latent individual or organization factors in the work environment contribute to active failures by clinicians on the sharp end and result in AEs unless adequate barriers or defenses are in place (Kreckler et al., 2008). The imperative to reduce errors has led to a multitude of safety interventions (Berwick, Calkins, McCannon, & Hackbarth, 2006; Hoffmann & Rohe, 2010). Strategies are listed in Tables 24.2 and 24.3. All these interventions may improve

Table 24.2 Patient safety systems interventions

<i>Method/approach</i>	<i>Description</i>	<i>Goal</i>
Incident reporting systems	• Identification, reporting, monitoring, and analysis	• Facilitate learning from errors and AEs to prevent similar events
Early warning scoring risk/severity scores	• Systems that assist staff to quickly identify patients at greatest risk of death at any moment, based on data from four physiological readings (e.g., systolic BP, HR, RR, and observation of level of consciousness)	• Prevent cardiac arrests and mortality in deteriorating patients through early recognition that facilitates fast responses to clinical instability
FMEA	• Careful review of steps in complex high-stakes clinical activities (for instance medication prescription and administration) to identify the points of greatest vulnerability	• Assist staff and leaders to optimize safety through a process that focuses their attention and identifies key priorities
Rapid response teams	• Team of clinicians bringing bedside critical care expertise to non-ICU hospitalized patients showing early signs of clinical deterioration	• Prevent cardiac or respiratory arrests and support clinicians/nurses working in direct patient care outside critical care settings to deliver optimal care
High-risk procedure SOPs	• Carefully drawn up lists of actions and considerations for routine use in settings such as surgical procedures (e.g. the WHO Surgical Safety Checklist)	• Prevent errors (e.g., wrong side procedures) • Reduce omissions and oversights
Information technologies	• Implementation of smartpumps, POC systems, etc.	• Reduce errors related to calculations and incorrect information related to doses or levels

Note: AE = adverse event; BP = blood pressure; FMEA = failure mode effect analysis; HR = heart rate; ICU = intensive care unit; POC = physician order computer entry; RR = respiratory rate; SOP = standard operating procedure; WHO = World Health Organization.

Table 24.3 Culture of safety and leadership interventions

<i>Method/approach</i>	<i>Description</i>	<i>Goal</i>
PDSA cycles	Structured but flexible system for developing, implementing, and evaluating changes in practice	Provide a manageable framework for front-line clinicians/managers to design initiatives to enhance logistics, safety, and quality
Team training Team coaching	Structured and semi-structured activities that raise staff awareness of pitfalls in communication and equip them with concrete strategies for working in teams	Reduce communication problems across members of health care teams that constitute a major source of errors and delays in care
Management walking rounds	Regularly planned visits of managers and executives to clinical settings to discuss safety and quality issues	Show organizational commitment to safety, build staff morale, and increase likelihood of workable, durable, properly resourced solutions
Staff surveys	Questionnaires that assess safety conditions, culture of safety, and leadership support for safety	Assess key underlying conditions influencing errors, pinpoint areas for improvement in specific settings
Structured communication	Rubrics/templates for information exchange (e.g., SBAR framework for exchanges between nurses, physicians)	Improve clarity, effectiveness of communication among providers

Note: PDSA = plan-do-study-act; SBAR = situation, background, assessment, recommendation.

patient safety to some extent. However, effectiveness varies from intervention to intervention and setting to setting, and the effect of some interventions have not been well studied. Taking the financial investments involved in these interventions into account (e.g., Pappas, 2008; see Uchida-Nakakoji & Stone, this volume) is infrequent, and documented benefits have not always been as impressive as expected. Moreover, sustained implementation and integration of safety-enhancing strategies is needed and, therefore, a system-based approach using bundled methods, rather than single interventions, is likely the most promising approach (Freeth et al., 2012; Hoffmann & Rohe, 2010).

When errors occur, effective response to patients and families are needed in the aftermath. Transparency – with straightforward disclosure, accompanied by emotional support – is increasingly favored to decrease patient and provider distress and avoid litigation (O'Connor, Coates, Yardley, & Wu, 2010). Anecdotal evidence suggests that forthrightness about errors and misadventures can strengthen patient–provider relationships without raising the overall costs of settling claims arising from actual or probable negligence. However, evidence and opinions remain mixed about the practicality or helpfulness to patients and families of disclosure of acts assumed to be errors under all circumstances (Bell et al., 2012).

Challenges

Patient safety research is challenging because methodological issues touch on almost all aspects of design, from data sources and defining events to attributing cause and evaluating safety improvement efforts. Variable risk of experiencing clinical events and poor outcomes in patients and

selection bias further complicate safety research. Ethical issues and sensitivities related to errors and AEs also create challenges.

Data sources and measurement of AEs and errors

Data for safety research are obtained from direct observation, administrative and clinical databases, and interviews/surveys. Each data source is potentially useful, but also associated with limitations (Table 24.4). Unobtrusive direct observation of care and/or clinical outcomes by trained research staff is theoretically best, but is only feasible in certain situations (e.g., hand hygiene, medication administration).

Consequently, information collected for administrative or clinical care purposes and stored in databases is often used. Databases containing hospital or clinic discharge summaries include information about patients, their care, and discharge disposition (patient returned home, was transferred to another facility, or died in hospital; Westra, Monsen, & Delaney, this volume). Additional information, such as clinical diagnoses, specific treatments, and secondary diagnoses for conditions that developed during care, may reveal risk factors for errors or AEs. In the case of secondary diagnoses, if time of onset has been recorded, preexisting conditions can be distinguished from complications of care. By making some assumptions, it is possible to estimate frequencies of AEs and errors using discharge data. Sometimes records can be reviewed to attempt to reconstruct care trajectories, understand how patient conditions changed over time, or both.

Table 24.4 Limitations of data sources for safety research

<i>Source</i>	<i>Limitation</i>
Direct observation	• Presence of an observer often impractical • Often time-consuming and costly • Ethical concerns of monitoring behavior
Clinical/administrative databases, patient records	• Permissions to use datasets or patient records are bound by legal and risk management considerations (chiefly around privacy) • Incomplete information on safety issues of interest • Inconsistent rates of reporting of incidents related low vs. high need for follow up • Inconsistent rates of documenting/reporting across different providers • Inconsistent reporting of incidents related to expectations that it is useful • Coding errors • Variable quality/thoroughness of records across facilities • Extraction of data often time-consuming, labor intensive, costly, and prone to errors
Interview/survey	• Subjects may have variable understandings of question or item content • Memory lapses and biases • Perceived risks of disclosing events or experiences may affect responses

Operational definitions for AEs and errors

Lack of reliable, valid and clear, uniformly applied definitions for AEs in general or for specific AEs (e.g., medication errors, nursing errors), and use of different measures and tools for extracting information from databases represent clear challenges for patient safety research in nursing (Burnett et al., 2013). Comparing results across studies is difficult, if not impossible, in the face of inconsistency in event definitions. In international research, ensuring relevance of indicators in various national contexts, careful forward and backward translation, and validation studies are essential. However, it appears as if there is a certain interoperability of operational definitions across health care systems. For instance, research findings suggest that the U.S.-developed Agency for Healthcare and Research and Quality (AHRQ) Patient Safety Indicator set of algorithms for use with discharge abstract data is usable across several countries (Drösler et al., 2009).

Retrospective review and attributions of cause

Estimating the frequency of safety problems, studying factors associated with error occurrences, and gauging intervention effects all require consistent identification of deviations from ideal practice and poor patient outcomes attributable to care. Usually, error events are not witnessed. AEs occurring subsequent to errors may take time to appear and when they do, their occurrence suggests, but cannot necessarily be unconditionally attributed to, errors. Patient care records are not dependable sources of information about errors because providers may not recognize their errors or, even if they realize they made errors, do not necessarily consistently document their acts. Differences in training and clinician personality may lead to variable attentiveness to certain types of errors or unintended consequences of clinical acts. There may also be disincentives to self-reporting errors, such as real or imagined fear of punishment.

Consequently, many statistics in the safety literature are likely underestimates of error frequencies, although it might be conjectured that errors with serious consequences are more consistently identified than errors that do not change patient clinical trajectories. All indirect means of detecting errors and AEs may overestimate or underestimate event occurrence. It is critical to recognize both possibilities. Sensitivity and specificity of methods for AE identification from patient records against a gold standard of direct observation of care or independent databases should be studied.

Patients have not usually been regarded as sources of information about errors. However, patients may be able to provide important perspectives on certain aspects of safety of health services (Doyle, Lennox, & Bell, 2013). Patients have been asked about their experiences of care and recall of certain elements of service (e.g., staff hand washing). Interviews and self-administered surveys may be used.

Staff have shown greater willingness to report undesirable events anonymously to researchers than through institutional reporting systems (Davis, Sevdalis, Neale, Massey, & Vincent, 2013). Therefore, under some circumstances, nurse-reported patient outcome data may be an alternative approach to obtaining prevalence or incidence data. For instance, nurses' long-term estimates of patient fall rates in surveys are concordant with continuously and systematically assessed fall occurrences (Cina-Tschumi, Schubert, Kressig, De Geest, & Schwendimann, 2009). In safety research, respondents are often asked whether they participate in rule breaking, deviate from professional practice standards, or hold opinions that run counter to those of their professional or institutional leaders. While such questions risk being seen as intrusive and potentially

self-incriminating, they have been included in studies at relatively low cost and have yielded important data.

Evaluating safety improvement efforts

Health care leaders are caught between knowing that safety problems are complex and multifactorial and, therefore, demanding complex interventions – and acknowledging real limitations in available resources to implement changes in practice or practice settings. Interventions to be tested should be scaled in relation to the severity of consequences of the errors or AEs. Very complex interventions aimed at lowering incidence of severe and complex problems should be “right sized” to a level that is sustainable over time.

Evaluation is the search for relevant information regarding the structures, processes, and outcomes of a service or program to permit decisions about continuing, expanding, modifying, or discontinuing it. Use of prospective data that addresses multiple components of the implementation process and track impacts from various perspectives (e.g., patient, clinician, management, health system) over time is preferred. Evaluation can involve qualitative and quantitative data regarding needs, structures, and immediate and downstream impacts of an intervention on the delivery of service.

Evaluating safety interventions and planning policy is complex. To understand the impact of an intervention, it can often be necessary to collect data not already being gathered, which may impose burdens on staff and can render evaluation plans unacceptable to those working in a setting. Shifts in safety indicators may not represent change that can be sustained in the short or long run without new investments of resources. Evaluating safety interventions can be costly, and external funding for such research may involve a high degree of competition. Conventional research funding competitions tend to value hypothesis-driven research; safety studies that test complex interventions with variable potential for implementation across clinicians and settings, take exploratory approaches or rely heavily on proxy measures – all compromises that can be necessary – which may not fare well in traditional peer review.

Variable risk of poor outcomes across recipients of care

Some outcomes are unacceptable regardless of patient characteristics or circumstances of care that influenced them (retained objects in surgical procedures, wrong patient medication administration errors, transfusion errors). Many types of events are not equally likely to occur in all patients, or do not reflect the same degree of provider and system failure. Simple counts of AEs and calculation of rates do not necessarily reflect quality of care. Comparison of AEs across units, clinics, institutions, or regions at one period of time, or in the same setting over time may result in the classic problem of comparing “apples to oranges” (Iezzoni, 1997) due to variation in case mix and procedures provided in different settings. Thus, error rates using denominators of total numbers of patients treated can be misleadingly inflated or deflated. Differences in rates where they are not expected or the absence of differences where one might expect to see them can be potentially explainable by different patient characteristics resulting in variations in baseline risk.

One solution is to calculate rates by patients sharing similar levels of risk (e.g., falls or pressure ulcers among “high-risk” patients). Another is to use denominators that reflect opportunities for error rather than patient volume. Rates can also be statistically adjusted for one or more patient characteristics, or individual patient risk can be estimated as a function of personal characteristics or numbers of treatments received (Iezzoni, 1997; Miller, Reardon, & Safi, 2001).

Combining data sources

Safety research often involves linking data from different sources – direct observation, surveys, record abstraction, use of secondary data. Matching patients across providers or institutions in databases of hospital discharge or clinic visit records can be technically complicated. Institutions or individual clinicians may appear under different names in different datasets and, on occasions, may be represented as collective entities (for instance, hospital systems rather than individual agencies). However, an even greater problem is that identifier data enabling linkages related to individual patients can raise important privacy concerns (for instance, social security or health insurance numbers).

Selection bias

Assembling samples that resemble the populations or conditions to which clinicians, managers, or policymakers would like to generalize is fundamental in all health research. Selection bias is a particular problem in safety research because of the sensitive issues involved. Organizations that refuse to contribute data, provide access to personnel, or otherwise participate in research studies will likely be systematically different from those whose leaders volunteer involvement. While it might be assumed that organizations with more developed safety and quality improvement programs and cultures would be more heavily represented in research studies, in practice it is difficult to tell if this is the case. Similarly, if risks of possible disclosure of potentially harmful information and the sensitivity of subject matter are added to the myriad factors influencing decisions of individuals to participate in studies, the possibility of biased subject pools even within organizations cannot be ruled out.

A multicenter approach is often needed to answer safety research questions involving variation in patient or organizational characteristics in relation to risk, to study effect sizes and differences that may be small, and to increase confidence in applicability and relevance of findings across settings. However, researchers using multicenter approaches must bear in mind that data are often collected in systematically different ways across institutions, and data that are not specifically gathered for research purposes will be particularly vulnerable to such variations.

Cross-national studies can offer interesting and useful perspectives, but involve particular challenges. Obtaining data can be very time-consuming if no unified local and/or national databases are available or are inconsistently available across sites. Terminology, variable definitions, and coding practices may vary greatly across and within countries (Krimsky et al., 2009). Careful attention to harmonizing operational definitions is essential to ensure that variations across countries represent true differences in patient safety variables, and are not attributable to variations in data quality and coding practices (Drösler et al., 2009).

Ethical challenges

Public confidence and trust are critical to smooth operation of health care services and livelihoods of providers. Not surprisingly, professionals have tended to believe that the public may somehow be better off not knowing about the failings of health services in general, and certainly not about imperfections of specific institutions or individual practitioners. A culture of silence about poor practices and bad outcomes has predominated. Discomfort with discussing safety issues has persisted even with demystification of the professions and the movement towards

transparency and public accountability in society and the rise of the patient safety movement. Even now, many are reluctant to admit that services may not be consistently of high quality for fear of jeopardizing reputations and exposing operations to external scrutiny.

Historically, errors were regarded as individual failings and system responses to error commonly involved punishment and reeducation intended to prevent future failures. The safety movement has been leading leaders away from assigning blame and towards considering environmental and system contexts for errors. Errors and omissions made in “good faith” are now distinguished from those made following blatant disregard for safety. Today, errors and near misses tend to be considered opportunities for learning, and policies that encourage secrecy are avoided. The shift from a blame culture to a just culture is not complete, however; clinicians in many institutions are still often reluctant to report errors or to have their colleagues learn of errors and incidents. In research studies, provisions for confidentiality can be implemented (and are often insisted upon by ethics committees), but the possibility of exposing one’s errors to supervisors, colleagues, or researchers still discourages participation in safety research (Krimsky et al., 2009).

Confidentiality, secure storage of data, and removal of patient and provider identifiers from datasets are all important in safety research. Unintentional disclosure of identifiable information related to clinical practice or critical incidents can have serious legal consequences, including civil or criminal liability claims. Inadvertently disclosed information about safety could threaten disciplinary action on clinicians’ licenses or threaten institutions’ accreditation status – especially if disclosed data were seen as reflecting systematic patterns of poor practice.

Many in health care today view reporting errors and near misses as a moral imperative, first to protect the immediately involved parties from unnecessary harm, and then to decrease risk for patients in the future by ensuring that useful information flows back into quality improvement processes locally and beyond. Given the short- and long-term stakes of errors in terms of patient suffering and mortality, it might also be considered a moral imperative to conduct research to reduce errors and accidents; by a similar token, failing to investigate safety problems and evaluate attempts to improve safety might be regarded as unethical.

Health professionals and researchers cannot ethically expose patients to risks unnecessarily for the sake of observing variation in practice (Johnstone & Kanitsaki, 2006). Special ethical considerations arise when observing actual clinical practice, if research staff are expected to be “nonreactive” (not to behave in ways that could change practice) while following higher-order ethical imperatives to act when there is clear evidence of potential preventable harm to a patient. Solutions include developing protocols for observers to handle situations where potentially serious problems in care emerge and using naïve (even nonclinician) research assistants for data collection. Naïve research assistants cannot usually recognize potential harm to a patient because they have been blinded to elements of the situation and/or their level of training and knowledge limits their awareness and, thus, may be relieved of a potential ethical conflict. However, researchers and review boards may not accept this approach. Anecdotally, researchers conducting fieldwork on medication safety have found ways to use clinicians as observers while respecting clinicians’ ethical imperatives to prevent harm by developing protocols where research staff pull aside a provider and unobtrusively point out that a serious error is about to be made (e.g., Nguyen, Nguyen, Haaijer-Ruskamp, & Taxis, 2014). Perhaps the most general statement that can be made is that ethical issues related to prospective observation of practice require forethought, and perhaps consultation with ethics specialists as well as clinical experts.

Future directions

Information technology

Many health care settings are either in the process of implementing or refining information technology (IT) solutions (Staggers & Nelson, this volume). Collaboration with IT specialists could provide researchers with better, more cost-effective ways of reliably capturing clinical observations, actions, and team functioning that could render safety research easier and less costly. Further, “smart” equipment now used in clinical settings can transmit or store data regarding settings, patient parameters received by the device, errors, alarms, and operator responses to them (e.g., Halpern, 2014). If privacy issues could be handled, such data could furnish important information about patient status, changes in patient status, and clinical decisions that would be useful alongside other data sources. To date, the potential of point-of-care technology for safety research has not been harnessed.

Simulation strategies

The relative infrequency of many patient safety problems makes them difficult to study prospectively using observational or experimental designs. Simulation has a long history in the health fields particularly for education (Rosen, 2008) and is an alternative approach for safety research especially around clinician behaviors, decision making, and teamwork. Studying practice phenomena (such as procedures or protocols) in laboratory or laboratory-like settings is an option worth considering because of potential advantages in terms of convenience and cost reductions in gathering data. Laboratory settings have the ethical benefits of not needing to expose patients to risk or burdening clinicians and care settings with data collection demands unnecessarily (Schmidt, Goldhaber-Fiebert, Ho, & McDonald, 2013).

Theory building

Themes based on accepted findings from the patient safety literature are listed in Table 24.5. Unifying multivariate, theoretical frameworks are needed to link relevant variables (including human factors and clinical aspects that interventions are designed to affect) with care performance (especially avoidance of errors) and patient outcomes, in context. In the face of finite and

Table 24.5 Themes in patient safety research findings

Theme

- Teamwork among health care workers and professionals is inconsistent across settings
 - Work environments, information sharing, and team elements are related to patient outcomes
 - Correlates of safety outcomes differ across contexts and studies and may operate differently across various clinical populations
 - Provider and environmental factors, and to a lesser extent patient factors, are modifiable elements influencing safety and are candidates for building interventions
 - Data about processes of care (observations, interactions, and actions) are generally more difficult to obtain than information about structures and outcomes of care
 - Research elucidating mechanisms that connect latent conditions to outcomes through processes of care has been very limited
-

dwindling resources for clinical care, it is likely to be extremely difficult to implement costly interventions that have unclear connections to downstream impacts when little if any evidence (or thought) exists about alternative means of arriving at those sought-after outcomes. Data from observational and experimental studies speaking to mechanisms of action for the underlying causes of errors will suggest where room exists for modifying them and assist in identifying intermediate endpoints.

Designs

Many forms of research are relevant to theory building for patient safety. Multilevel modeling (Grittner & Lahmann, this volume), structural equation modeling (Schmiege & Bryan, this volume), analysis of longitudinal data to track changes in sets of variables over time (Deboeck & Boker, this volume; Stoddard, this volume) will all generate data relevant to hypothesized linkages. When intervention research touches on safety considerations, random assignment (Rubin & Zell, this volume) can be difficult to justify to prospective subjects (whether patients, clinicians or health care settings). Thus, “natural experiments,” such as staggered rollout of safety initiatives of various types across institutions or clinical areas, can create “control” conditions useful for assessing the impacts of the initiatives can be optimized using principles of quasi-experimental design (Eckardt & Rindskopf, this volume). To do this, a special challenge lies in recognizing the impending rollout of a project or program and obtaining baseline measures of conditions, processes of care, or outcomes in all settings before the initiatives are launched in any of the areas.

Multivariate perspectives

There has now been extensive study of individual factors or small sets of variables in safety research. Because it is intuitively obvious and increasingly supported by empirical data that multiple factors come together to influence patient safety, it is important to develop interventions that are appropriately complex and theoretically driven. Interventions that address multiple levels and engage multiple actors in patient safety situations simultaneously would appear to have the best likelihood of actually improving outcomes.

Interventions having multiple components and bundling of interventions are both a practical and ethical necessity. Where it is really necessary to tease out differential impacts of different components because time and money restrict how many components can be implemented at once, factorial designs (comparing the effects of different combinations of interventions or different composition of interventions in terms of bundled components) may be especially useful. Trials and studies of information technology, equipment, and procedure changes should be enhanced with concurrent measurement of individual patient and clinician variables, and consideration of impacts on workflow and teamwork to enrich explanatory and implementation potential of the data that emerges. Use of multivariate, multilevel designs will accelerate development of evidence-based management approaches, particularly in terms of tailoring interventions to the demands and characteristics of settings and making well-informed decisions about tradeoffs when not all can be done.

Safety-related interventions and workload

One safety issue is usually researched at a time. In clinical settings, single interventions or small bundles are evaluated to the exclusion of others. In practice, multiple simultaneous demands for change (initiative overload, initiative fatigue) can create considerable job stress, undermining

success of initiatives, as well as negatively affecting patient outcomes. A sense of the critical mass of change that individuals and organizations can tolerate, as well as the larger question of exactly where, in the face of increasing demands on health care systems by clients and stagnant or shrinking health care system budgets, nurses and other professionals should best direct their energy is urgently needed. Rich practical and theoretical ground remains to be mined by investigating cost-benefit tradeoffs associated with workload, prioritization in care, and the human factors and workload issues generated by change. To date, the proliferation of safety-oriented interventions and new technologies has generated remarkably little research or debate.

Conclusion

Safety research in general and nursing-focused and nurse-led research on safety issues has generated important information about AEs and the nature of the underlying multidisciplinary clinical concerns. A number of interventions appear to reduce rates of some types of AEs, particularly in acute care and medical surgical settings. The ideas and approaches in safety research apply to a wide range of situations, including community-based and long-term care, where well-chosen care delivered attentively and in a timely manner influence patient outcomes. Safety research is still dominated by descriptions of problems and correlates of poor outcomes. Development of theory and theoretically grounded interventions lags. To reflect the complexity of the health care environment, advances will rely on application of sophisticated research methods and models.

Individuals and societies are paying a great deal for care that is not always effective and is not particularly consistent or safe. The question is where to go about improving things. Nurses are commonly thought of as being at the emotional and ethical heart of health care – the nobility of their intentions has rarely been questioned. In the new data-based era in health care – with transparency in health care being at a premium and scarcity of resources an increasing feature of policy debates at all levels – the question is clear: Will nurses be perceived as barriers to achieving safer health care, or will their willingness to rethink care delivery supported by close study of problems, contexts, and solutions herald a better future for health care systems? Progress in safety research in nursing thus far suggests much promise for improving delivery of care in fast-changing and resource-constrained settings.

References

- Aiken, L. H., Clarke, S. P., Sloane, D. M., Sochalski, J., & Silber, J. H. (2002). Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *JAMA* 288, 1987–1993.
- Balas, M. C., Scott, L. D., & Rogers, A. E. (2004). The prevalence and nature of errors and near errors reported by hospital staff nurses. *Applied Nursing Research*, 17, 224–230.
- Bell, S. K., Smulowitz, P. B., Woodward, A. C., Mello, M. M., Duva, A. M., Boothman, R. C., & Sands, K. (2012). Disclosure, apology, and offer programs: Stakeholders' views of barriers to and strategies for broad implementation. *Milbank Quarterly*, 90, 682–705.
- Berwick, D. M., Calkins, D. R., McCannon, C. J., & Hackbarth, A. D. (2006). The 100,000 lives campaign: Setting a goal and a deadline for improving health care quality. *JAMA*, 295, 324–327.
- Burnett, S., Renz, A., Wiig, S., Fernandes, A., Weggelaar, A. M., Calltorp, J., . . . Fulop, N. (2013). Prospects for comparing European hospitals in terms of quality and safety: Lessons from a comparative study in five countries. *International Journal for Quality in Health Care*, 25, 1–7.
- Cho, S.-H., & Yun, S.-C. (2009). Bed-to-nurse ratios, provision of basic nursing care, and in-hospital and 30-day mortality among acute stroke patients admitted to an intensive care unit: Cross-sectional analysis of survey and administrative data. *International Journal of Nursing Studies*, 46, 1092–1101.
- Cina-Tschumi, B., Schubert, M., Kressig, R. W., De Geest, S., & Schwendimann, R. (2009). Frequencies of falls in Swiss hospitals: Concordance between nurses' estimates and fall incident reports. *International Journal of Nursing Studies*, 46, 164–171.

- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.
- Curry, L. A., Nembhard, I. M., & Bradley, E. H. (2009). Qualitative and mixed methods provide unique contributions to outcomes research. *Circulation*, 119, 1442–1452.
- Davis, R. E., Sevdalis, N., Neale, G., Massey, R., & Vincent, C. A. (2013). Hospital patients' reports of medical errors and undesirable events in their health care. *Journal of Evaluation in Clinical Practice*, 19, 875–881.
- De Meester, K., Van Bogaert, P., Clarke, S. P., & Bossaert, L. (2013). In-hospital mortality after serious adverse events on medical and surgical nursing units: A mixed methods study. *Journal of Clinical Nursing*, 22, 2308–2317.
- Diya, L., Van den Heede, K., Sermeus, W., & Lesaffre, E. (2012). The relationship between in-hospital mortality, readmission into the intensive care nursing unit and/or operating theatre and nurse staffing levels. *Journal of Advanced Nursing*, 68, 1073–1081.
- Doyle, C., Lennox, L., & Bell, D. (2013). A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open*, 3, e001570.
- Dröslér, S. E., Klazinga, N. S., Romano, P. S., Tancredi, D. J., Gogorcena Aoi, M. A., Hewitt, M. C., . . . Kelley, E. (2009). Application of patient safety indicators internationally: A pilot study among seven countries. *International Journal for Quality in Health Care*, 21, 272–278.
- Freeth, D., Sandall, J., Allan, T., Warburton, F., Berridge, E. J., Mackintosh, N., . . . Abbott, S. (2012). A methodological study to compare survey-based and observation-based evaluations of organisational and safety cultures and then compare both approaches with markers of the quality of care. *Health Technology Assessment*, 16, i–iv, 1–184.
- Halpern, N. A. (2014). Innovative designs for the smart ICU. Part 2: The ICU. *Chest*, 145, 646–658.
- Harris, D. M., Westfall, J. M., Fernald, D. H., Duclos, C. W., West, D. R., Niebauer, L., . . . Main, D. S. (2005). Mixed methods analysis of medical error event reports: A report from the ASIPS collaborative. In K. Henriksen, J. B. Battles, E. S. Marks, & D. I. Lewin (Eds.), *Advances in patient safety: From research to implementation: Vol. 2. Concepts and methodology*. Rockville, MD: Agency for Healthcare Research and Quality.
- Henriksen, K., & Kaplan, H. (2003). Hindsight bias, outcome knowledge and adaptive learning. *Quality and Safety in Health Care*, 12, ii46–ii50.
- Hoffmann, B., & Rohe, J. (2010). Patient safety and error management: What causes adverse events and how can they be prevented? *Deutsches Ärzteblatt International*, 107, 92–99.
- Iezzoni, L. I. (1997). *Risk adjustment for measuring healthcare outcomes* (2nd ed.). Chicago, IL: Health Administration Press.
- Institute of Medicine. (2000). *To err is human: Building a safer health system*. Washington, DC: National Academies Press.
- Institute of Medicine. (2004). *Keeping patients safe: Transforming the work environment of nurses*. Washington, DC: National Academies Press.
- Johnstone, M.-J., & Kanitsaki, O. (2006). The ethics and practical importance of defining, distinguishing and disclosing nursing errors: A discussion paper. *International Journal of Nursing Studies*, 43, 367–376.
- Kale, A., Keohane, C. A., Maviglia, S., Gandhi, T. K., & Poon, E. G. (2012). Adverse drug events caused by serious medication administration errors. *BMJ Quality & Safety*, 21, 933–938.
- Kane, R. L., Shamliyan, T. A., Mueller, C., Duval, S., & Wilt, T. J. (2007). The association of registered nurse staffing levels and patient outcomes: Systematic review and meta-analysis. *Medical Care*, 45, 1195–1204.
- Kreckler, S., Catchpole, K., Bottomley, M., Handa, A., & McCulloch, P. (2008). Interruptions during drug rounds: An observational study. *British Journal of Nursing*, 17, 1326–1330.
- Krimsky, W. S., Mroz, I. B., McIlwaine, J. K., Surgenor, S. D., Christian, D., Corwin, H. L., . . . Malayaman, N. (2009). A model for increasing patient safety in the intensive care unit: Increasing the implementation rates of proven safety measures. *Quality and Safety in Health Care*, 18, 74–80.
- Lawton, R., McEachan, R. R. C., Giles, S. J., Sirriyeh, R., Watt, I. S., & Wright, J. (2012). Development of an evidence-based framework of factors contributing to patient safety incidents in hospital settings: A systematic review. *BMJ Quality & Safety*, 21, 369–380.
- Mardon, R. E., Khanna, K., Sorra, J., Dyer, N., & Famolaro, T. (2010). Exploring relationships between hospital patient safety culture and adverse events. *Journal of Patient Safety*, 6, 226–232.
- Merwin, E., & Thornlow, D. (2006). Methodologies used in nursing research designed to improve patient safety. *Annual Review of Nursing Research*, 24, 273–292.
- Meurier, C. E. (2000). Understanding the nature of errors in nursing: Using a model to analyse critical incident reports of errors which had resulted in an adverse or potentially adverse event. *Journal of Advanced Nursing*, 32, 202–207.

- Miller, C. C., Reardon, M. J., & Safi, H. J. (2001). *Risk stratification: A practical guide for clinicians*. Cambridge: Cambridge University Press.
- Nguyen, H.-T., Nguyen, T.-D., Haaijer-Ruskamp, F.M., & Taxis, K. (2014). Errors in preparation and administration of insulin in two urban Vietnamese hospitals: An observational study. *Nursing Research*, 63, 68–72.
- O'Connor, E., Coates, H. M., Yardley, I. E., & Wu, A. W. (2010). Disclosure of patient safety incidents: A comprehensive review. *International Journal of Quality in Health Care*, 22, 371–379.
- Pappas, S. H. (2008). The cost of nurse-sensitive adverse events. *Journal of Nursing Administration*, 38, 230–236.
- Patrician, P. A., Loan, L., McCarthy, M., Fridman, M., Donaldson, N., Bingham, M., & Brosch, L. R. (2011). The association of shift-level nurse staffing with adverse patient events. *Journal of Nursing Administration*, 41, 64–70.
- Reason, J. (1990). *Human error*. Cambridge: Cambridge University Press.
- Richardson, A., & Storr, J. (2010). Patient safety: A literature review on the impact of nursing empowerment, leadership and collaboration. *International Nursing Review*, 57, 12–21.
- Robert, G. B., Anderson, J. E., Burnett, S. J., Aase, K., Andersson-Gare, B., Bal, R., . . . QUASER team. (2011). A longitudinal, multi-level comparative study of quality and safety in European hospitals: The QUASER study protocol. *BMC Health Services Research*, 11, 285.
- Rosen, K. R. (2008). The history of medical simulation. *Journal of Critical Care*, 23, 157–166.
- Schmidt, E. M., Goldhaber-Fiebert, S. N., Ho, L. A., & McDonald, K. M. (2013). Use of simulation exercises in patient safety efforts. In P. G. Shekelle et al. (Eds.), *Making health care safer II. An updated analysis of the evidence for patient safety practices* (pp. 439–460). Bethesda, MD: Agency for Healthcare Research and Quality.
- Singer, S., Lin, S., Falwell, A., Gaba, D., & Baker, L. (2009). Relationship of safety climate and safety performance in hospitals. *Health Services Research*, 44, 399–421.
- Twigg, D., Duffield, C., Bremner, A., Rapley, P., & Finn, J. (2011). The impact of the nursing hours per patient day (NHPPD) staffing method on patient outcomes: A retrospective analysis of patient and staffing data. *International Journal of Nursing Studies*, 48, 540–548.
- Vincent, C. (2010). *Patient safety* (2nd ed.). Hoboken, NJ: Wiley-Blackwell.
- Vogus, T. J., & Sutcliffe, K. M. (2007). The impact of safety organizing, trusted leadership, and care pathways on reported medication errors in hospital nursing units. *Medical Care*, 45, 997–1002.
- von Laue, N. C., Schwappach, D. L., & Koeck, C. M. (2003). The epidemiology of medical errors: A review of the literature. *Wiener Klinische Wochenschrift*, 115, 318–325.
- Wagner, C., Smits, M., Sorra, J., & Huang, C. C. (2013). Assessing patient safety culture in hospitals across countries. *International Journal for Quality in Health Care*, 25, 213–221.
- Weissman, J. S., Rothschild, J. M., Bendavid, E., Sprivulis, P., Cook, E. F., Evans, R. S., . . . Bates, D. W. (2007). Hospital workload and adverse events. *Medical Care*, 45, 448–455.

ECONOMIC EVALUATIONS FOR NURSING RESEARCH

Mayuko Uchida-Nakakoji and Patricia W. Stone

Healthcare decisions are complex and healthcare is expensive. Unlike most daily choices, decisions made in the clinical setting have substantial consequences and involve uncertainties and tradeoffs that need to be carefully weighed. Treatment benefit versus risk, survival versus quality of life, and cost versus health improvement are just few of the many difficult decisions clinicians encounter daily. Decision making in healthcare is based on several factors such as information and resources available to patients and clinicians, as well as personal preferences and costs (Hunink et al., 2001; Spetz, 2005). Therefore, we need to know not only the efficacy of an intervention but also the attributable costs.

Attention to healthcare costs has grown over time and there is increased demand to utilize economic evaluations to inform the comparative cost-effectiveness of clinical decisions in global and national settings. For example, in 2014, major coverage expansions from the Affordable Care Act will begin and U.S. health spending growth is estimated to rise to 7.4% from the projected average growth rate of 4% for 2011 to 2013. By 2021, the healthcare share of the gross national product is estimated to reach nearly 20% from its 2010 level of 17.9% (Keehan et al., 2012). This increasing level of healthcare expenditures has drawn attention from patients, providers, institutions, and especially policymakers who now expect economic outcomes as part of evidence for evaluating the effectiveness of new treatments, interventions, and health technologies. Unsustainable healthcare spending is not the only contributing factor that has drawn attention to costs. Today, there is growing consensus among policymakers that part of the solution to improve healthcare efficiently and effectively, is to shift the way clinical research is conducted.

Recognizing the need to improve care while curbing growth of healthcare spending, Congress passed two recent federal initiatives including: (a) the American Recovery and Reinvestment Act (the stimulus spending); and (b) the Patient Protection and Affordable Care Act (U.S. national health reform). The American Recovery and Reinvestment Act authorized \$1.1 billion more to be spent on comparative effectiveness research (CER) and designated the money to three agencies: National Institutes of Health (NIH), the Agency for Healthcare Research and Quality (AHRQ), and the Office of the Secretary of Health and Human Services. As part of the national health reform legislation, a new, non-governmental entity called the Patient-Centered Outcomes Research Institute (PCORI) was created; this institute was established to oversee and set guidelines for CER. The law also created a steady stream of research funding for PCORI. Starting in 2013, Medicare and all private health insurance companies will pay a tax into a trust fund that will support the activities of PCORI. This funding is estimated to reach \$500 million annually

by 2015. Furthermore, PCORI will contract with NIH, AHRQ, and others to oversee funding and research initiatives. The purpose of CER is to contribute to both clinicians' and patients' decision-making processes and while the health care reform legislation did not include language advocating for formal cost-effectiveness analyses, it has been acknowledged by many national thought leaders that economic considerations are an important aspect of any comprehensive comparative effectiveness assessment (Garber, 2011).

Over the last several years, numerous articles have appeared in the literature examining economic evidence of nursing care (Dawes et al., 2007; Ganz, Simmons, & Schnelle, 2005; Goodman et al., 2008; Postmus et al., 2011; Sorensen & Frich, 2008). Cost-effectiveness analysis is one of the fastest growing fields in health services research; yet various terminologies, steps to conduct and interpret its findings make this type of evaluation ambiguous for nurse investigators. Several economic methods exist to address economic outcomes of healthcare interventions. This chapter intends to: (a) provide an overview of the basic types of economic evaluations, (b) present outcome measures for each type of analyses, (c) provide steps for analyses, and (d) discuss challenges and issues to consider when conducting these evaluations.

Types of economic evaluations

Economic evaluations are fundamentally about making choices (Stone, Curran, & Bakken, 2002). In general, all economic analyses compare two or more alternative courses of action. Five methods of economic evaluations are frequently reported in the literature (see Table 25.1). In all of these evaluations, alternative strategies are compared and the incremental costs and effectiveness of the competing strategies are examined. These methods differ only in how the effectiveness (benefits or consequences) of interventions are valued. The focus remains on incremental changes in effectiveness (i.e., comparing the outcome of one intervention with that of another) (Stone, 2001a). Some basic terminologies that come up during these evaluations are outlined in Table 25.2. Overviews of the five methods of evaluation follow.

Table 25.1 Types of economic evaluations

<i>Type of Study</i>	<i>Definition</i>	<i>Measurement/ Valuation of Costs</i>	<i>Identification of Consequences</i>	<i>Measurement/ Valuation of Consequences</i>
Cost Minimization Analysis (CMA)	An analysis that calculates the incremental costs of alternatives that achieve similar health outcomes	Monetary Units	None	None
Cost Consequence Analysis (CCA)	An analysis in which costs and benefits are listed without any attempt to aggregate them	Monetary Units	Single or multiple effects, not necessarily common to both alternatives	Naturally occurring units (e.g., readmission time, disease cases avoided)
Cost-Benefit Analysis (CBA)	An analysis in which incremental costs and benefits are computed and all benefits and costs are measured in monetary units	Monetary Units	Single or multiple effects, not necessarily common to both alternatives	Monetary units

Cost-Effectiveness Analysis (CEA)	An analysis in which incremental costs and benefits are presented in an incremental cost-effectiveness ratio (ICER)	Monetary Units	Single effect of interest, common to both alternatives but achieved to different degrees	Naturally occurring units (e.g., life-years gained, disability-days saved, disease cases avoided)
Cost-Utility Analysis (CUA)	A special type of cost-effectiveness analysis that considers quality of life	Monetary Units	Single or multiple effects, not necessarily common to both alternatives	Healthy years (e.g., QALYs)

Table 25.2 Common terminology in economic evaluations

<i>Term</i>	<i>Definition</i>
Boundaries of the study	The scope of the study
Competing alternative	An alternative strategy for diagnosing or treating the disease or health condition (comparator)
Consumer Price Index (CPI)	A measure of average changes in price over time; used to adjust costs that are estimated in different years
Decision analysis model	A model used to calculate expected values of health strategies; most frequently used means for calculating incremental cost-effectiveness ratios
Discounting	Process of converting future costs and effects into present terms; the discount rate used in cost-effectiveness analysis is 3% for both costs and quality-adjusted life years
Health intervention	A strategy for diagnosing or treating a disease or health condition
Health states	Health status (condition) of an individual or population at one point in time (e.g., alive, well, dead)
Incremental cost-effectiveness ratio	Ratio of the difference of the costs of two alternatives to the difference in effectiveness between the same two alternatives; calculated in cost-effectiveness and cost-utility analyses
Net benefit	Calculated by the Total Benefits minus Total Costs
Perspective	Viewpoint from which the analysis is conducted
Preference score	A weighted value for life lived in a given health state ranging from 0 to 1, where 0 represents death and 1 represents perfect health
Quality Adjusted Life Year (QALY)	A year of life lived in perfect health
Sensitivity analysis	An analysis in which a parameter is varied in the model to examine how they might vary in alternative circumstances
Probabilistic sensitivity analysis	Calculations in which one or more parameters are varied randomly, using a defined distribution to determine the proportion of times that an intervention is economically favored in comparison with other interventions
Time horizon	The period of time over which all costs and effects are measured in an economic analysis
Utility	Preference or desirability to be in a certain health state

Cost-minimization analysis

If the benefits or consequences of the competing alternatives under consideration are assumed to be similar but the costs are different, one can simply compare the costs by conducting an economic evaluation called cost-minimization analysis (CMA). The goal of CMA is to choose the intervention with the lowest cost (minimal cost strategy). Often these studies begin as larger evaluation studies in which the investigators expected one intervention to be both more effective and more expensive. Consider the study by Goodman and colleagues (2008). Clinical outcome measures were collected prior to the study being published as a CMA and the authors measured numerous outcomes of a fitness-for-life program and found no statistically significant differences between groups on costs or clinical outcomes. Another CMA study (Schuurman et al., 2009) examined the costs for prevention and treatment of pressure ulcers to identify the least resource-intensive pressure ulcer prevention strategy (human versus technical approach). The investigators collected several cost outcome measures during a larger cohort study and ultimately found that the incident rate did not differ between the two strategies, but the technical approach was less expensive.

While CMA is relatively straightforward, it is rarely an appropriate method of analysis because health outcomes are oftentimes *not* considered having equivalent effectiveness. Furthermore, individual preferences and other important values that go into making healthcare decisions are not accounted for which limits CMA from being considered a full economic evaluation. Of course, this method can be used whenever two or more interventions of equal effectiveness are being compared; however, one should remember that this method alone does not answer whether the intervention is cost-effective.

Cost-consequence analysis

As noted, in most cases, the benefits or consequences across health interventions are not equal. In such a case, a cost-consequence analysis (CCA) can be carried out in which decision makers can first comprehensively list the consequences of the competing alternative strategies and then determine the relative values of these outcomes. Those costs and consequences of two or more alternatives are measured, but listed separately. This approach is often selected when there is no clear summary measure to explain the outcomes of the interventions being studied. Because there is no single measure to evaluate the effectiveness of an intervention, a CCA does not allow direct comparison of the outcomes across studies. In a CCA the investigators are expected to formulate their own opinions about the relative value of the findings by listing all the various consequences applicable to each strategy. Two studies serve as examples of CCA in the nursing literature. Sorensen and Frich (2008) analyzed the consequences of a nurse follow-up intervention for chronic non-malignant pain patients and described outcomes using the SF-36 health status questionnaire subscales. Dawes et al. (2007) compared a specialist nurse supported discharge program versus usual care for women receiving major abdominal or pelvic surgery. The authors reported SF-36 survey responses, surgery complications, length of hospital stay, readmissions, costs, and satisfaction of women at discharge.

Cost-benefit analysis

Decisions about healthcare services involve competing choices between discrete interventions. The most simple decision rule would be to choose the intervention with the greatest net benefits (effectiveness minus cost). This type of comparison in which the benefits or consequences are only measured monetarily is called a cost-benefit analysis (CBA). CBA is a broader form of

economic evaluation and the final result is always reported as a monetary value of net benefits (e.g., dollars) or in the form of a cost to benefit ratio. The primary goal of CBA is to identify interventions where the net benefit is greater than zero, which suggests that the value of the benefits is greater than the costs. It is a useful technique for allocation efficiency under a fixed budget where one can rank interventions according to their net benefits. Most recently, a post-discharge care transition program among Medicare beneficiaries was examined in a randomized controlled trial using CBA (Saleh, Freire, Morris-Dickinson, & Shannon, 2012). While the authors found no significant differences in readmission rates between the two groups, they calculated a cost-to-benefit ratio of the discharge program to be 1.09; meaning that for every dollar spent on the program, a savings of \$1.09 was realized from reduced readmission frequency. The major disadvantage or difficulty that arises with CBA is the notion of attaching a monetary value on health outcomes. Many decision makers find this idea uncomfortable and even unethical; Is it right to put a monetary value on human life and suffering? Such ethical issues on placing a dollar value to life have encouraged investigators to look to other forms of economic evaluation (Stone, 2001a).

Cost-effectiveness analysis

By way of definition, cost-effectiveness analysis (CEA) is a form of economic evaluation where both the incremental costs and consequences are examined (Drummond, Sculpher, Torrance, O'Brien, & Stoddart, 2005). Typically in a CEA the incremental consequences are measured in a single common unit, such as life-years gained or adverse cases avoided. A cardinal feature of CEA is how the effects are summarized in an incremental cost-effectiveness ratio (ICER), which is defined as

$$\text{ICER} = (C_1 - C_2) / (E_1 - E_2) \quad (25.1)$$

where C_1 equals the cost of the new intervention (often newer and more expensive), C_2 equals the cost of the comparator (often the standard of care), E_1 equals the effect of the new intervention, and E_2 equals the effect of the comparator. The ICER provides information on how much one needs to spend to realize a unit gain in effectiveness.

Decision trees

In a CEA, analysts often use data collected through the process of a randomized controlled trial often powered on an outcome other than the cost-effectiveness result (e.g., Postmus & colleagues (2011) initially set out to examine heart failure readmission rates of two nurse-led disease management programs), or more frequently utilize a decision analytic approach and model the problem by using a decision tree (Kang, Mandsager, Biddle, & Weber, 2012). A sample decision tree is shown in Figure 25.1. A decision tree is a mathematical modeling technique used to calculate the costs and effects associated with events in an event pathway (Detsky, Naglie, Krahn, Redelmeier, & Naimark, 1997). In the figure, the decision is between choosing alternative 1 or alternative 2. Both alternatives have associated probabilities of good and bad outcomes. Additionally, there are associated costs of each strategy. In the CEA by Kang and colleagues, investigators employed a decision tree model to evaluate the cost-effectiveness of three alternative screening strategies for methicillin-resistant *Staphylococcus aureus* (MRSA) among patients admitted to hospitals. For each of the three screening strategies, the authors indicated the pathways to where the patient was screened, the screening result, and the occurrence of MRSA infection. They also calculated the costs associated with each strategy and displayed their results using ICERs to compare the additional costs to prevent a MRSA infection.

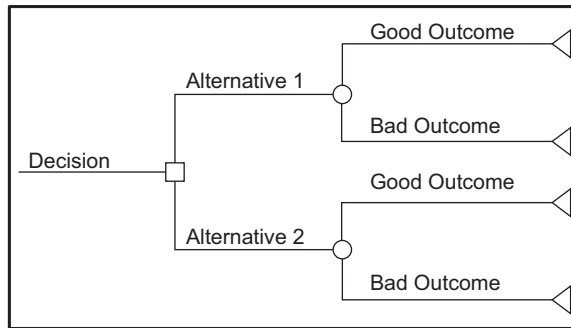


Figure 25.1 Sample decision tree.

Markov cohort simulation model

Other analytic approaches used in CEA include a Markov cohort simulation model (Sonnenberg & Beck, 1993). In this form, a simulation is carried out to see what may happen to a cohort of individuals over multiple periods of time under the same trial. Compared to a decision tree where there is a finite time horizon and individuals can only receive an intervention once, individuals in a Markov simulation model are allowed to transition through the model more than once and capture recursive events. An example of this method is seen in the study by Ganz, Simmons, and Schnelle (2005). The authors estimated the incremental cost-effectiveness of recommended nurse staffing levels versus median staffing levels in nursing homes. In this analysis the patients were hypothetically assigned to one of the two staffing scenarios. The authors created a Markov model that represented the typical flow of patients from the first day in the nursing home to discharge, with a probability of being hospitalized for five major health conditions. Other outcomes of interest in this study were life expectancy, and quality-adjusted life years, which leads directly into the more detailed discussion of what it means to use health effects in terms of quality-adjusted life expectancy and their applications.

Cost-utility analysis

Although the terms are used interchangeably, health economists often distinguish cost-utility analysis (CUA) from CEA. From a practical standpoint, a CUA can be considered a special type of CEA; what makes CUA unique is that this type of evaluation is conducted when individual preferences are taken into account and quality of life (i.e., the utility) is measured. Although there is general consensus and gradual consistency in published reports, many studies in health sciences, especially those studies conducted in the United States refer to CUA as a CEA even though utility measurements are used to value effectiveness.

It is important to remember that a CUA examines intervention benefits by integrating both the quantity (for example, years of life gained by intervention) and quality of life in a single multidimensional measure called quality adjusted life years (QALYs). The main purpose of QALYs are to (a) provide a comparable measure across many interventions or programs, (b) incorporate individual values or preferences for different health states, and (c) integrate morbidity and mortality into a single measure.

Quality of life is accounted for in the form of a utility, which is a measure of preference or desirability for a given health state (i.e., health status or condition). A utility is rated on a scale of 0 (death) to 1 (perfect health). This resulting value of utilities is called a preference score. For instance, individuals who consider themselves to be in perfect health would rate their quality of

life as a 1, and individuals who are close to dying would rate their life as a 0; on the other hand, individuals who are living with a chronic debilitating disease might rate their life as worth only sixth-tenth of a year of life lived in perfect health. In that case, for these individuals, one year of their QALY for one year would be calculated as follows: $0.6 \times 1 \text{ year} = 0.6 \text{ QALYs}$.

Several methods exist for obtaining preference scores to calculate QALYs such as the standard gamble, time-trade-off or rating scores. More detail about using these methods is provided in CEA textbooks (Drummond et al., 2005; Muennig, 2008) and interested readers are encouraged to seek further training. Other utility assessment instruments such as the EuroQOL 5 item scale (www.euroqol.org/home.html) or the more commonly known SF-36 survey scores (www.sf-36.org/) are also used to elicit preference scores or QALY weights.

Since dollars spent to gain a QALY are not disease specific, the measure is useful for informing health policy decisions and is recommended for use by the U.S. Public Health Service's Panel on Cost-Effectiveness in Health and Medicine (Gold, Siegel, Russell, & Weinstein, 1996; Weinstein, 1996). While there is no universal agreement as to what society should be willing to pay to gain a QALY, the figure of \$50,000/QALY is often cited and used to evaluate healthcare interventions. While this numerical willingness to pay threshold may depend on the perspective (i.e., viewpoint) of the analysis or type of disease, in general it translates to: for 1QALY gained, decision makers are willing to pay up to \$50,000.

Other types of CEAs, besides CUA, have been conducted when the clinical outcome measures are numerical such as lives saved from the intervention, mmHg of blood pressure reduced, disease cases averted. In a CUA, however, multiple outcomes can be incorporated and the outcomes are not just counted but are valued according to their desirability (utility).

There are many CUAs in the peer-reviewed literature. For example, one group of researchers (Pham et al., 2011) compared current standard of care with four pressure ulcer prevention strategies in nursing homes: (a) use of pressure redistribution mattresses, (b) oral nutritional supplements, (c) skin emollients, and (d) foam cleansers during incontinence care. They examined lifetime risks and costs of developing pressure ulcers by characterizing the effectiveness of the various strategies as cost/QALYs gained. While all strategies produced QALY gains, the authors found that using pressure redistribution mattresses and foam cleansers compared to the current prevention practice produced more benefits at lower mean lifetime costs than providing nutritional supplements which was \$7.8 million per QALY gained. For interested readers, additional CUA study examples are available on the CEA Registry (<https://research.tufts-nemc.org/cear4/>).

Conducting economic evaluations

This chapter only provides a basic overview of how to begin to conduct these analyses and should be interpreted as one approach to starting an economic evaluation. The general steps in conducting economic evaluations are provided in Figure 25.2. Specialized training and collaboration with health economists is recommended. Excellent resources are available in forms of readings (Drummond et al., 2005; Hunink et al., 2001; Muennig, 2008; Stone, 2001a, 2001b) and online modules (www.herc.research.va.gov/training/training_course_cea.asp).

Choosing the appropriate economic evaluation: study question and study type

Before the appropriate economic evaluation method can be selected, it is important to identify a key question or problem that needs to be solved. This is a paramount step in choosing the appropriate evaluation method, and can also be the hardest. The challenge is in narrowing down the research question to a well-defined statement, to clearly indicate the population of interest and to

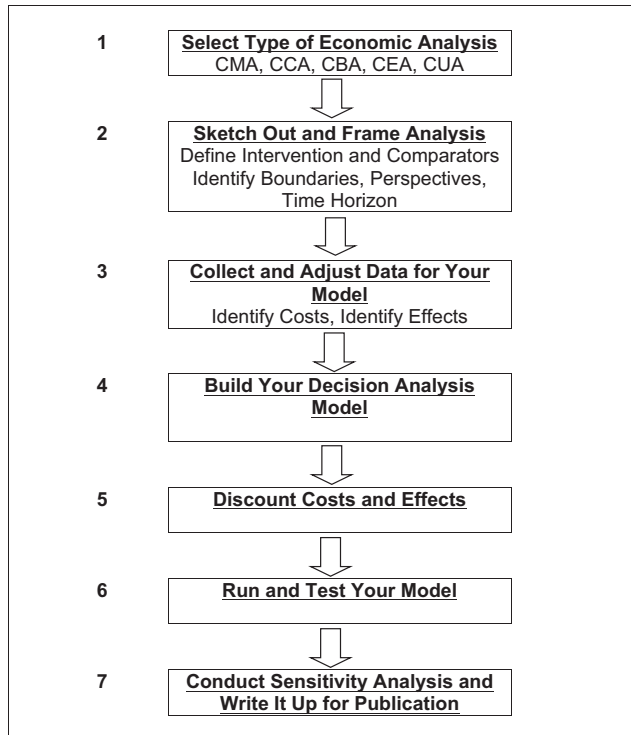


Figure 25.2 Basic steps in approaching economic evaluations.

define the interventions and comparators. Additional considerations in selecting the type of analysis include: (a) goals of the analysis (who is the audience for the analysis); (b) whether the effectiveness of the interventions are similar or equal (in which case a CMA can be done); (c) the effectiveness measures available (e.g., can QALYs be generated); (d) potential impact of the interventions on either quality or quantity of life (if quality of life is important, then a CUA may be appropriate); (e) availability of the data sources; (f) analytic expertise available; and (g) ethical issues.

Sketching and framing the analysis

Once the economic method has been selected, the analysis should be sketched out and framed. In designing an economic evaluation, the investigator should thoroughly review the literature for background information on health conditions and interventions under evaluation. By taking this step, the researcher will have a better understanding of the length of time that will be modeled, the various outcomes that need to be considered, and the appropriate comparator(s). Often, new interventions are compared against the standard of care or usual practice. Standard of care can be a useful comparator because investigators can provide policymakers with concrete information about the cost-effectiveness of the program by comparing it to the current baseline.

Establishing the *boundaries* (or scope) of the study early in the analysis is important because this allows one to limit the costs and effects to be included in the study. Many interventions have some degree of spillover effects. Therefore it is important to draw practical and feasible limits around the analysis.

In all types of economic evaluations, the *perspective* or viewpoint taken in the analysis can also drive the set of costs and benefits to be considered. For example when conducting a CEA, a hospital

may only be concerned with how much money it will have to spend and how much it might save on treatment or prevention strategies. It is very unlikely that the hospital will be interested in how much money the patient has to spend on transportation to get to the hospital for a preventive intervention. On the other hand, if the goal of the analysis affects broad resource allocation and applies to society as a whole, then the societal perspective is appropriate and recommended (Gold et al., 1996). The societal perspective, which is the broadest one and always relevant, incorporates all costs and all health effects regardless of who incurs them. A rule of thumb when selecting a perspective is to take on the societal perspective first and then, if necessary, present the same results from a different perspective. Using the societal perspective is also recommended for a reference case analysis. The idea for a reference case analysis developed out of need to standardize methodological approaches and reporting of CEA studies so results from across different studies could be compared.

The analytical *time horizon* is the period over which all costs and effects are considered in the analysis. Time horizons may vary from the duration of a hospital stay to the patient's entire life span (to death). The appropriate time horizon to consider depends on the intervention and the effects being studied. Once the framing of the analysis is conducted, the investigator is now ready to gather data to list costs and effects.

Costs and effects

Costs

Obviously, costs are important elements of economic evaluations. There are several methods for measuring and valuing costs. Data relevant to the analysis may come from published studies, electronic databases, and other sources such as medical or financial experts. As mentioned, the particular range of costs to be used for the analysis depends on the perspective of the study and the type of interventions employed. First, it is necessary to identify the amount of resources attributed or consumed under a given strategy. One should outline the categories of costs included using the list in Table 25.3. The costs of medical visits, hospitalizations, and lab services are called direct costs. Opportunity costs or costs accrued indirectly that do not include the consumption of goods and services can include time spent away from work due to a medical appointment. While opportunity costs can still be monetized (calculate time spent away from work based on wage), there are other costs that cannot be easily accounted for. Such costs are called intangible costs that have no perceived market value such as pain and suffering or the monetary value of year of healthy life. In a CEA, intangible costs are not monetized but are captured in the denominator of the cost-effectiveness ratio (Drummond et al., 2005; Muennig, 2008).

Once the resource costs are categorized, monetary valuations of resource costs are estimated. This process of estimating costs is called costing. Costing can take considerable time and effort; it is important for the investigator to determine how precise the cost estimates need to be in a given analysis that determines whether one conducts gross costing (using gross costs from a

Table 25.3 Basic cost elements of economic evaluations

Type of Costs	Definition
Direct Medical Costs	All costs incurred by medical care providers when treating a condition
Direct Non-Medical Costs	Costs imposed on non-medical care personnel, including patients and their caregivers
Opportunity Costs	Indirect cost of the time influenced by the illness such as lost productivity due to illness, injury, or premature death

single data source) or micro-costing (accounting for every resource used). When cost estimations are derived from various sources, it is important to standardize all costs to the same currency and year. For instance, Pham and colleagues (2011) expressed all monetary values in 2009 Canadian dollars even though they used cost estimates from studies conducted in 1994 and 2003. This is because prices have inflated since 1994, and therefore the investigator can no longer just use cost estimates from those years. When older cost data are used, they may underestimate the cost of medical care in current practice. Because a dollar in 2009 does not have the same purchasing power as a dollar in 2014, the costs from different years need to be adjusted or inflated to a standard year format. This can be done using the Consumer Price Index (CPI). To inflate older cost data in the United States, one can visit the Bureau of Labor Statistics (BLS) website (www.bls.gov/cpi/home.htm). The BLS also calculates a medical inflation rate. Because medical costs are rising more rapidly than costs in other markets, health economists often recommend using the medical inflation rate to inflate costs that pertain only to health care resources.

Probabilities

Probabilities are another important element of economic evaluations. Probability values for events being modeled conditional on the treatments being analyzed are necessary for constructing the decision analytic model shown in Figure 25.1. For instance, to calculate the cost of having the bad outcome under alternative 1, one will need to know the likelihood of the bad outcome event occurring under alternative 1. For CEA studies under a randomized controlled trial or cohort study, primary probabilities are used; however, when evaluations are based on secondary data sources, probability values are often times retrieved from the literature. Once the relevant probabilities are identified, the investigator can calculate the path probabilities for each intervention strategy.

Discounting costs and effects

Once all costs and benefits have been calculated, future costs and benefits need to be discounted to the net present value. This is based on the notion that in general, people place more weight for immediate gains over future gains. For example if someone offered \$100 today versus \$103 a year from now, one would probably forgo the extra \$3 to receive the money today. Discounting is a process to account for time preferences and measures how much a dollar payable in the future is worth today. Most economic evaluations include future costs. For instance, if a treatment for diabetes prevents hospitalizations that occur one year or more into the future, these costs need to be discounted. In economic evaluations, future costs and benefits (DFCB) are discounted using the following formula:

$$\text{DFCB} = CF / (1 + r)^n, \quad (25.2)$$

where CF = cost of future event (usually measured in dollars at today's value), r = the discount rate, and n = number of years in the future (Drummond et al., 2005; Muennig, 2008). Currently, in the United States as set forth by the U.S. Public Health Service's Panel on Cost-Effectiveness in Health and Medicine, both costs and effects should be discounted using a 3% discount rate (Gold et al., 1996).

Running the analysis

Conducting these economic evaluations by hand can be done, but is often not recommended for obvious reasons. Not all of these methods require computer software assistance but in a typical CEA or CUA where multiple factors are compared simultaneously, calculation errors can be minimized

through the use of software programs such as Microsoft Excel and TreeAge (TreeAge Software; Williamstown, MA). While different software packages exist to aid with analyses, TreeAge is one of the widely used programs by healthcare analysts because they have packages specifically designed to address healthcare utilization. TreeAge is also used to teach decision analytic techniques in classroom settings because of its user-friendly interface and readily available student licenses.

Synthesizing results

Once the analysis is complete, the next important step is to consider what the results mean. What does 1 QALY gained for \$100,000 spent mean? For \$10,000? Are these results applicable to the local clinical setting? Why or why not? Such questions like this may arise because results gained from economic evaluations need to be synthesized and translated to patients, other health professionals, and especially policymakers. There is often the misconception surrounding the term “cost-effective” because there is variable and loose use of the term. The confusion is inherent in assuming that “cost-effective = cost-saving.” In reality, this is not the case. For instance, an intervention that costs \$100,000 for each QALY it produces relative to the next most effective alternative may be viewed as expensive by some, but might be purchased by others. Just because something is not cost-saving, it does not necessarily mean that it is not an effective and efficient option. It is imperative that we see “cost-effective” as having the additional benefit worth the additional cost.

Sensitivity analysis

As seen in the way data were derived in decision analytic models, many of the values collected for economic analyses (inputs) include a certain degree of uncertainty. Uncertainty in this context applies to instances when the optimal decision remains unclear due to lack of clearly supported evidence. Therefore it would be flawed to simply assume that the result from a single analysis is absolute. Running alternative analyses by varying the parameters of interest are called *sensitivity analyses*. These analyses indicate the degree of influence the particular value has on the analysis and should be conducted to explore stability of estimates under various assumptions (i.e., changing the range of parameter estimates). There are several different modalities of sensitivity analyses. First, one way (also called univariate) sensitivity analysis allows the investigator to vary a parameter of interest to see how changing the values for a single parameter may influence the study results. When the investigator varies two variables over their range of plausible values while holding all else constant, this is called a two-way (bivariate) sensitivity analysis. Ganz et al. (2005) used a series of one-way sensitivity analyses to explore which parameter led to the greatest change in the incremental cost-effectiveness of increasing nurse staffing levels.

Although one-way sensitivity analyses are one of the most common forms of sensitivity analyses in the literature, varying one source of uncertainty at a time is usually inadequate. In such a case, multivariate sensitivity analyses can be carried out by examining multiple sources of uncertainty simultaneously. Additionally, an investigator can carry out what is called a *probabilistic sensitivity analysis* (Doubilet, Begg, Weinstein, Braun, & McNeil, 1985). In this form of sensitivity analysis, probability distributions are applied to the specific ranges for the various parameters and samples are drawn at random from these distributions. Kang and colleagues (2012) conducted a probabilistic sensitivity analysis using Monte Carlo simulations for 1,000 cycles. In this process, a hypothetical cohort of patients was entered into the decision analytic model to encounter a number of different probabilities such as the likelihood of being screened for MRSA. The authors found that even in their 1,000 trial simulation, the targeted screening strategy that they found to

be cost-effective prior to running the sensitivity analysis was still dominant, meaning the intervention cost less while preventing more MRSA infections.

Conclusions

While there may be no one right answer, at the end of the day, healthcare decisions need to be made and must be made with evidence to support the outcomes. Because economic evaluations can be conducted in many subtly different forms it is up to the investigator to clearly identify the appropriate analytic approach and use standardized methods set forth by the Panel (Weinstein, 1996). The Consolidated Health Economic Evaluation Reporting Standards (CHEERS) checklist in Table 25.4 may be helpful in assessing economic evaluations in healthcare, as well as useful for those conducting and publishing these types of analyses (Husereau et al., 2013).

Becoming familiar with basic economic evaluation terminologies and techniques is an important first step in taking part in these decision-making efforts. Economic evaluations are a critical part of any well-functioning society; however, it should be noted that a CEA is merely an aid to decision making and is not a complete procedure for making resource allocation decisions in healthcare. As more clinicians and researchers actively take on the role of informing and shaping health policy decisions, it is important that they continue to generate economic evidence for the care they provide to demonstrate the cost-effectiveness of care.

Table 25.4 CHEERS checklist: Items to include when reporting economic evaluations of health interventions

Section/item	Recommendation	Page
Title and abstract		
1-Title	Identify study as an economic evaluation, or use specific terms such as “cost-effectiveness analysis” and describe interventions compared.	
2-Abstract	Provide structured summary of objectives, perspective, setting, methods (include design and inputs), results (include base-case and uncertainty analyses), and conclusion.	
Introduction		
3-Background and objectives	Provide explicit statement of broader context for study. Present study question and relevance for health policy or practice decisions.	
Methods		
4-Target population and subgroups	Describe characteristics of base-case population and subgroups analyzed; say why they were chosen.	
5-Setting and location	State relevant aspects of the system(s) in which the decision(s) need(s) to be made.	
6-Perspective	Describe perspective and relate this to the costs being evaluated.	
7-Comparators	Describe interventions or strategies being compared; state why they were chosen.	
8-Time horizon	State time horizon(s) over which costs and consequences are being evaluated; say why appropriate.	
9-Discount rate	Report choice of discount rate(s) used for costs and outcomes; say why appropriate.	
10-Health outcomes	Describe outcomes used as measure(s) of benefit and relevance for type of analysis performed.	
11-Effectiveness measurement	a-Single study-based estimates: Describe design features fully; describe why the single study was a sufficient source of clinical effectiveness data. b-Synthesis-based estimates: Describe fully methods used to identify included studies and synthesis of clinical effectiveness data.	
12-Preference-based outcomes	If applicable, describe the population and methods used to elicit preferences for outcomes.	

Section/item	Recommendation	Page
13-Resources and costs	a-Single study-based outcome evaluation: Describe approaches used to estimate resource use associated with the alternative interventions. Describe primary or secondary methods for valuing each resource item in terms of its unit cost. Describe any adjustments made to approximate to opportunity costs. b-Model-based economic evaluation: describe approaches and data sources used to estimate resource use associated with model health states. Describe primary or secondary research methods for valuing each resource item in terms of its unit cost. Describe any adjustments made to approximate to opportunity costs.	
14-Currency, price date, and conversion	Report dates of the estimated resource quantities and unit costs. Describe methods for adjusting estimated unit costs to the year of reported costs, if necessary. Describe methods for converting costs into a common currency base and the exchange rate.	
15-Model	Describe and give reasons for the specific type of decision-analytic model used. Providing a figure to show model structure is strongly recommended.	
16-Assumptions	Describe all structural or other assumptions underpinning the decision-analytic model.	
17-Analytic methods	Describe all analytic methods supporting the evaluation. This could include methods for dealing with skewed, missing, or censored data; extrapolation methods; methods for pooling data; approaches used to validate or make adjustments (e.g., half-cycle corrections) to a model; and methods for handling population heterogeneity and uncertainty.	
Results		
18-Parameters	Report values, ranges, references, and if used, probability distributions for all parameters. Report reasons or sources for distributions used to represent uncertainty where appropriate. Providing a table to show input values is strongly recommended.	
19-Incremental costs and outcomes	For each intervention, report mean values for the main categories of estimated costs and outcomes of interest, as well as mean differences between comparator groups. If applicable, report incremental cost-effectiveness ratios.	
20-Uncertainty	a Single study-based economic evaluation: Describe effects of sampling uncertainty for estimated incremental cost, incremental effectiveness, and incremental cost-effectiveness, together with impact of methodological assumptions (like discount rate, perspective) b Model-based economic evaluation: Describe effects on results of uncertainty for all input parameters, and uncertainty related to model structure and assumptions.	
21-Heterogeneity	If applicable, report differences in costs, outcomes, or cost-effectiveness that can be explained by variations among patient subgroups with different baseline characteristics, or other observed variability in effects that are not reducible by more information.	
Discussion		
22-Findings, limitations, generalizability, & current knowledge	Summarize key findings. Describe how findings support the conclusions reached. Discuss limitations and the generalizability of findings, and how findings fit with current knowledge.	
Other		
23-Funding sources	Identify funding source, and state funder role in identification, design, conduct, and reporting of analyses. Describe nonmonetary sources of support.	
24-COIs	Report any potential for COIs among study contributors, following journal policy. If no journal policy, recommendation is to follow ICMJE COI policy.	

Note: CHEERS = Consolidated Health Economic Evaluation Reporting Standards; COI = conflicts of interest; ICMJE = International Committee of Medical Journal Editors. From Husereau et al. (2013). This article was published in *Value in Health*, 16, Husereau et al., Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement, e1-35. Copyright Elsevier, 2013. Used with permission.

References

- Dawes, H. A., Docherty, T., Traynor, I., Gilmore, D. H., Jardine, A. G., & Knill-Jones, R. (2007). Specialist nurse supported discharge in gynaecology: A randomised comparison and economic evaluation. *European Journal of Obstetrics, Gynecology, & Reproductive Biology*, 130, 262–270.
- Detsky, A. S., Naglie, G., Krahn, M. D., Redelmeier, D. A., & Naimark, D. (1997). Primer on medical decision analysis: Part 2 – Building a tree. *Medical Decision Making*, 17, 126–135.
- Doubilet, P., Begg, C. B., Weinstein, M. C., Braun, P., & McNeil, B. J. (1985). Probabilistic sensitivity analysis using Monte Carlo simulation. A practical approach. *Medical Decision Making*, 5, 157–177.
- Drummond, M. F., Sculpher, M., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes* (3rd ed.). Oxford: Oxford University Press.
- Ganz, D. A., Simmons, S. F., & Schnelle, J. F. (2005). Cost-effectiveness of recommended nurse staffing levels for short-stay skilled nursing facility patients. *BMC Health Services Research*, 5, 35.
- Garber, A. M. (2011). How the Patient-Centered Outcomes Research Institute can best influence real-world health care decision making. *Health Affairs*, 30, 2243–2251.
- Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (Eds.). (1996). *Cost-effectiveness in health and medicine. Report of the Panel on Cost-Effectiveness in Health and Medicine*. New York, NY: Oxford University Press.
- Goodman, H., Parsons, A., Davison, J., Preedy, M., Peters, E., Shuldham, C., . . . Cowie, M. R. (2008). A randomised controlled trial to evaluate a nurse-led programme of support and lifestyle management for patients awaiting cardiac surgery 'Fit for surgery: Fit for life' study. *European Journal of Cardiovascular Nursing*, 7, 189–195.
- Husereau, D., Drummond, M., Petrou, S., Carswell, C., Moher, D., Greenberg, D., . . . Force, C. T. (2013). Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *Value in Health*, 16(2), e1–e5.
- Hunink, M. G. M., Glasziou, P., Siegel, J. E., Weeks, J. C., Pliskin, J. S., Elstein, A. S., & Weinstein, M. C. (2001). *Decision making in health and medicine: Integrating evidence and values*. Cambridge: Cambridge University Press.
- Kang, J., Mandsager, P., Biddle, A. K., & Weber, D. J. (2012). Cost-effectiveness analysis of active surveillance screening for methicillin-resistant *Staphylococcus aureus* in an academic hospital setting. *Infection Control & Hospital Epidemiology*, 33, 477–486.
- Keehan, S., Cuckler, G. A., Sisko, A., Madison, A. J., Smith, S., Lizonitz, J., . . . Wolfe, C. (2012). National health expenditure projections: Modest annual growth until coverage expands and economic growth accelerates. *Health Affairs*, 31, 1600–1612.
- Muennig, P. (2008). *Cost-effectiveness analysis in health: A practical approach* (2nd ed.). San Francisco, CA: Jossey-Bass.
- Pham, B., Stern, A., Chen, W., Sander, B., John-Baptiste, A., Thein, H.-H., . . . Krahn, M. (2011). Preventing pressure ulcers in long-term care: A cost-effectiveness analysis. *Archives of Internal Medicine*, 171, 1839–1847.
- Postmus, D., Pari, A. A., Jaarsma, T., Luttik, M. L., van Veldhuisen, D. J., Hillege, H. L., & Buskens, E. (2011). A trial-based economic evaluation of 2 nurse-led disease management programs in heart failure. *American Heart Journal*, 162, 1096–1104.
- Saleh, S. S., Freire, C., Morris-Dickinson, G., & Shannon, T. (2012). An effectiveness and cost-benefit analysis of a hospital-based discharge transition program for elderly Medicare recipients. *Journal of the American Geriatrics Society*, 60, 1051–1056.
- Schuurman, J.-P., Schoonhoven, L., Defloor, T., van Engelshoven, I., van Ramshorst, B., & Buskens, E. (2009). Economic evaluation of pressure ulcer care: A cost minimization analysis of preventive strategies. *Nursing Economics*, 27, 390–400.
- Sonnenberg, F. A., & Beck, J. R. (1993). Markov models in medical decision making: A practical guide. *Medical Decision Making*, 13, 322–338.
- Sorensen, J., & Frich, L. (2008). Home visits by specially trained nurses after discharge from multi-disciplinary pain care: A cost consequence analysis based on a randomised controlled trial. *European Journal of Pain*, 12, 164–171.
- Spetz, J. (2005). The cost and cost-effectiveness of nursing services in health care. *Nursing Outlook*, 53, 305–309.
- Stone, P. W. (2001a). Dollars and sense: A primer for the novice in economic analyses (Part I). *Applied Nursing Research*, 14, 54–55.
- Stone, P. W. (2001b). Dollars and sense: A primer for the novice in economic analyses (Part II). *Applied Nursing Research*, 14, 110–112.
- Stone, P. W., Curran, C. R., & Bakken, S. (2002). Economic evidence for evidence-based practice. *Journal of Nursing Scholarship*, 34, 277–282.
- Weinstein, M. C. (1996). Recommendations of the Panel on Cost-effectiveness in Health and Medicine. *JAMA*, 276, 1253–1258.

26

MIXED METHODS

Louise Doyle

Mixed methods research is the combination of quantitative and qualitative approaches within a study for the broad purpose of obtaining breadth and depth (Johnson, Onwuegbuzie, & Turner, 2007; Tashakkori & Creswell, 2007). Over the past two decades, mixed methods research has seen a significant increase within health and nursing research. This chapter provides an overview of how qualitative and quantitative methods can be combined to better meet the aims and objectives of a research study. The focus is to highlight what mixed methods research can do and to identify some mixed methods designs that enrich research findings.

Emergence of mixed methods

The methodology utilized by nurse researchers has undergone many changes since the era of Florence Nightingale and her research on the Crimean War (Parahoo, 2006; Polit & Beck, 2006). Nursing research at that time and for many decades to follow was dominated by the positivist paradigm and associated quantitative approaches. The quantitative approach was suited to the study of interventions and outcomes, but sole utilization of quantitative methods encouraged a reductionist approach to research where aspects of the person and their health and illness experiences were studied. An understanding of this limitation sparked intense interest in using other approaches to study human phenomena (Streubert Speziale & Carpenter, 2007).

In the 1970s and 1980s, qualitative research emerged in nursing with the realization that qualitative methods had the potential to consider the complex and multifaceted constructs that could not be addressed using quantitative methods solely. Qualitative researchers believed that adopting these approaches allowed for a new perspective with potential to create greater depth of understanding of people's thinking, feeling, behavior, and motivations. The movement towards qualitative research as a method to better understand a person's experience was congruent with the holistic belief system at the center of professional nursing practice (Streubert Speziale, & Rinaldi Carpenter, 2007). Subsequent explosion in the popularity of qualitative methods was reflected in the number of nursing research studies which were grounded in naturalistic enquiry (Beck, 2013; Parahoo, 2006; Streubert Speziale, & Rinaldi Carpenter, 2007).

Paradigm wars ensued. The relativist ontology, subjectivist epistemology, and naturalistic procedures that compose the constructivist foundation of qualitative research (Denzin & Lincoln, 2005) were seen as antithetical to the philosophical assumptions of positivism and post-positivism.

The resultant 'incompatibility thesis' held that research methods from the quantitative and qualitative perspectives could never be mixed (Johnson & Onwuegbuzie, 2004), and the argument of incommensurability had a captive audience for a significant period. Recently, increasing frustration with methodological purism and paradigm purists (Rallis & Rossman, 2003) paved the way for quantitative and qualitative research methods to be used in the same study (Greene, 2007; Howe, 1988; Johnson & Onwuegbuzie, 2004; Morgan, 2007; Onwuegbuzie, 2002). Critical theory, the dialectic stance, critical realism, and the transformative paradigm were identified as suitable underlying philosophies (Teddle & Tashakkori, 2010). However, pragmatism is most widely used (Johnson & Gray, 2010), and is serving as a rapprochement to the opposite sides of the paradigm wars.

Pragmatism began as a philosophical movement of American origin during the latter 19th century. The philosophy of pragmatism advanced the notion that consequences are more important than process (Doyle, Brady, & Byrne, 2009). It rejects traditional dualisms and seeks instead to find a middle ground between philosophical dogmatisms (Johnson & Onwuegbuzie, 2004). Pragmatism calls for a convergence of qualitative and quantitative methods when necessary and appropriate (Feilzer, 2010). Thus, pragmatism allows selection and justification of research methods based on the premise that they are best suited to answer research questions. Any approach to research that contributes to an understanding of the phenomenon under investigation is seen as worthy.

Mixed methods is now one of the fastest growing approaches in research design (Bergman, 2008). Use of mixed methods within nursing research has increased significantly (Andrew & Halcomb, 2009). Mixed methods designs are particularly suited for use in social, behavioral, and health research where understanding of complex and multifaceted phenomena is required (Twinn, 2003). An example is a study by Ossebaard, Seydel, and van Gemert-Pijnen (2012) examining usability issues of a national online health resource among patients with long-term conditions and their carers. Mixed methods with a variety of data collection techniques including standardized self-report scales, individual and focus group interviews, and completion of online scenarios were used to develop an understanding of usability issues that could not have been accessed using solely one research method.

In nursing science, mixed methods facilitates objective study of nursing outcomes and enables understanding of the human experience as well. Mixed methods can facilitate a more person-centered approach to nursing research with emphasis on the individual, *in addition* to focus on the variables (cf., Henly, 2007). The sections that follow detail the rationale for selecting a mixed methods research design and outline how mixed methods can address a range of research questions.

Rationale for mixed methods designs

Justification of the mixed methods approach is essential. This section outlines the most commonly identified rationales for undertaking a mixed methods study. In other words it identifies what mixed methods research can actually do. The overarching rationale for use of a mixed methods design is that it is needed to answer the research question, for example when both confirmation and exploration are required (Teddle & Tashakkori, 2009). The following list presents an amalgamation of detailed and nuanced justifications identified by a number of authors (Bryman, 2006; Creswell & Plano Clark, 2007; Greene, Caracelli, & Graham, 1989).

- Answer different research questions: A research problem generates a number of questions that cannot be answered using quantitative or qualitative approaches alone. Both are therefore required to meet study objectives. This often occurs in nursing studies, where complex and multifaceted questions are commonly posed.

- Explanation of findings: A quantitative element may provide data requiring further explanation and interpretation. This information can be obtained using a qualitative component whereby participants explain their quantitative responses. In evaluation studies, survey administration may be followed by qualitative interviews (e.g., Bamberger, Rao, & Woolcock, 2010), or semi-structured interviews may be incorporated into surveys (Schatz, 2012). In development of item banks for patient reported outcomes, cognitive interviews (Beatty & Willis, 2007) are used to obtain respondent feedback about language, ambiguity, comprehensibility, and relevance of the question (DeWalt, Rothrock, Yount, & Stone on behalf of the PROMIS Cooperative Group, 2007).
- A requirement for exploration first: A qualitative phase may be required to develop hypotheses which then require testing quantitatively. Similarly, qualitative interviews can be conducted to identify relevant variables that can then be studied quantitatively.
- Offsetting weaknesses and providing stronger conclusions: A mixed methods design may be utilized to investigate a research question from multiple perspectives, thereby allowing the limitations of each approach to be neutralized while strengths are built upon (Bryman 2006; Doyle et al., 2009). This provides stronger conclusions than may be achieved using a single-approach design.
- Triangulation: In a triangulation study the aim is to study the same phenomena using both quantitative and qualitative methods. Convergence of findings increases the validity of the findings. In addition, divergence of findings can provide greater insight into complex aspects of a phenomenon (Teddlie & Tashakkori, 2011).
- Meeting funding requirements: Although a practical rather than a methodological issue, the availability of research funding may be a consideration. In a fiscal environment with limited funding, sponsors seeking more value for their money may prefer studies with mixed methods designs (Creswell, 2011; Giddings, 2006). However, there is a danger that researchers driven by funder preferences will propose mixed methods designs in a tokenistic manner without adequate consideration of suitability from a methodological perspective.
- Enhancement of findings: A common rationale for conducting a mixed methods study is to enhance findings, the logic being that when findings from a quantitative design can be further enhanced by findings from a qualitative design (and vice versa) a mixed methods approach is best suited. Increasingly in nursing research, this rationale is provided to support the embedding of qualitative methods into intervention studies.

Intervention research

Findings from randomized controlled trials (RCTs) provide high-level evidence about intervention efficacy because design properties support inference about cause, relative to a comparison treatment (Rubin & Zell, this volume; West & Thoemmes, 2010). When RCTs are not feasible for practical or ethical reasons, alternative quasi-experimental designs are used in intervention research (Eckardt & Rindskopf, this volume; West et al., 2008). Therapeutic processes and mechanisms producing change are not always explored in an RCT, which may limit application of findings (Campbell et al., 2007). Similarly, it would be insufficient to discover that an intervention does not work without information about why or how it might be improved. Generalizability, or effectiveness of the intervention in real-world settings, clinical significance, and importance to patients and families are key concerns (Flemming, 2007; Relton, Torgerson, O’Cathain, & Nicholl, 2010; Song, Sandelowski, & Happ, 2010). Further, acceptability of an intervention is critical to its use in practice (Gorad, 2010; Marcus, Stuart, Wang, Shadish, &

Steiner, 2012). RCTs are used to estimate the population average treatment effect (Rubin & Zell, this volume), which may not be relevant to specific individuals who are non-responders. Incorporating qualitative approaches into intervention studies using mixed methods can play an important role in developing a larger information base from which to interpret the statistical findings in intervention research (Muncey, 2009; Sandelowski, 1996). Interviews, direct observation, and focus groups may be embedded into RCTs and quasi-experimental studies to provide this information.

Qualitative methods may be incorporated at many stages, determined by the purpose of their inclusion in the first instance (Creswell, Fetters, Plano Clark, & Morales, 2009). Before the trial begins, qualitative methods may be used to develop an instrument or to inform recruitment. During the trial, qualitative data may be used to understand the impact of the intervention on participants, to understand participant experiences during the trial, or to identify potential mediating and moderating factors. After the trial, qualitative methods can inform a more in-depth understanding of how the intervention worked (or did not work), how the treatment could be improved, or what the meaning of the change was for participants. Mixed methods research affords the opportunity to explore statistically non-significant results or extreme cases who responded particularly well or not at all. Sandelowski (1996) described how a qualitative component can be used to study within- and between-subject variations on an outcome. If an intervention proved beneficial, yet some individuals in the experimental group showed no improvement or even deteriorated, qualitative methods can facilitate scrutiny of these variations. Qualitative methods may also be used to study cases in the control group who deviate from the norms of their group.

Quantitative researchers electing to utilize qualitative approaches in their intervention studies need to be cautious in their approach. Song et al. (2010) noted that in many intervention studies utilizing a qualitative component, the qualitative work is often complementary rather than integral to the trial work and is tacked on with seemingly little forethought. Indeed, Creswell et al. (2009) cautioned against simply “adding on” a qualitative element to an otherwise quantitative study and instead advise that the language and philosophy of mixed methods research be used to guide the trial from the start. Keeping in mind these caveats however, it appears that mixed methods intervention studies have much to offer nursing research by helping to preserve the clinical relevance of these studies by placing findings in a clinical context from both patient and healthcare professional perspectives (Flemming, 2007).

Choosing a mixed methods design

Mixed methods research continues to evolve and this is to some degree evidenced by the fact that there is as yet no absolute agreed upon, exhaustive typology of mixed methods research designs (Teddle & Tashakkori, 2009). Instead, a number of basic and complex typologies with a number of commonalities but also significant differences have been proposed (Greene & Caracelli, 1997; Creswell, Plano Clark, Gutmann, & Hanson, 2003; Teddle & Tashakkori, 2006; Creswell & Plano Clark, 2007; Leech & Onwuegbuzie, 2009; Teddle & Tashakkori, 2009; Creswell & Plano Clark, 2011). The increasing taxonomy of mixed methods designs presents researchers with a variety of options, the choice of which is dependent on the nature of the research question and the degree of integration required. It is beyond the remit of this chapter to provide a comprehensive account of each typology. Instead, an overview of the typology developed by Creswell & Plano Clark (2007) will be provided. Although later expanded upon to identify six rather than four major mixed methods designs (Creswell & Plano Clark, 2011), the original typology is presented here as it is an accessible one when considering the *basic* type of mixed methods design required

for a research study. Once a general research design has been chosen, referral back to a number of the texts identified previously is required for a more detailed discussion of the chosen design and its variants.

Selection of a particular design is largely guided by methodological decisions that a researcher must consider; three major considerations (Creswell & Plano Clark, 2007) are:

- The timing of the quantitative and qualitative methods: whether the stages need to be done concurrently or sequentially. This is generally determined by the nature of the research question and the rationale for collecting each dataset.
- The weighting of the quantitative and qualitative methods: whether both methods are assigned equal priority or whether one approach is more fundamental to the research question than the other. Generally, exploratory mixed methods research assigns priority to the qualitative phase while explanatory designs assign priority to the quantitative phase. However, there are exceptions to this general rule and many exploratory and explanatory studies assign equal priority to the quantitative and qualitative elements of the study.
- The mixing of the quantitative and qualitative methods: where in the project lifetime the mixing of quantitative and qualitative approaches will occur. Integration may occur at a number of points including the development of research questions, data collection, data analysis, and, importantly, the final interpretation/discussion phase of a study.

Consideration of these three key issues provides novice mixed methods researchers with a strong foothold into interpreting the different designs available (Creswell, 2011). Following a consideration of these issues, a mixed method design may be selected from among four major types (Creswell & Plano Clark, 2007). These are the triangulation design (identified as the convergent design in Creswell & Plano Clark, 2011), the embedded design, the explanatory design, and the exploratory design. An overview of these four designs is provided here with reference to studies that have utilized these designs to illustrate their real-world research problems.

Triangulation (convergent) design

The triangulation design is primarily used when a research aim involves convergence. Using triangulation, quantitative findings can be confirmed or corroborated with qualitative findings (Creswell & Plano Clark, 2007). The triangulation design is concurrent as quantitative and qualitative data are gathered at the same time. Data are generally analyzed separately and results are merged in the interpretation phase of the study, where qualitative and quantitative findings are given equal weight. The desired endpoint of a triangulation study is arrival at valid and well-substantiated conclusions about a single phenomenon (Creswell & Plano Clark, 2007). The biggest challenge facing researchers who use this design is the decision about what to do if there is little or no convergence between the quantitative and qualitative findings. Additional data may be needed to more fully investigate the divergence (Creswell & Plano Clark, 2007) or the discrepant findings may be further explored (Teddle & Taskakori, 2010). Vaughan Dickson, Lee, and Riegel (2011) used a concurrent triangulation mixed methods design to explore how knowledge and cognitive functioning influence heart failure self-care. Their main purpose in using this design was to strengthen validity of findings by examining concordance by comparing qualitative evidence of self-care and knowledge with the results derived from quantitative measures of self-care, and identify cases where there was inconsistency.

Embedded design

The embedded design is characterized by having one dominant method, with the other data set providing a supportive role (Doyle et al., 2009). The methods may be conducted concurrently or sequentially. This design is chosen when a qualitative or quantitative method alone would be insufficient to answer all of the research questions. The most common type of embedded design is one in which a qualitative component is embedded within a larger quantitative design as is the case with the embedded experimental design (Creswell & Plano Clark, 2007). In this design, the qualitative data can inform the development of a treatment and can also be utilized post-intervention to elaborate on the treatment results, including vital information about why the intervention worked or did not work (Flemming, 2007). An example of the benefits of an embedded design can be seen in McCabe, Roche, Hegarty, & McCann (2011). This research used an RCT to determine the effects of new media art using a virtual window on health-related quality of life in patients experiencing stem cell transplantation compared to a treatment as usual control group. Quantitative findings showed that those in the intervention group had lower scores on measures of depression and anxiety, on average, and overall had better experiences. Findings from the embedded qualitative element explained why better outcomes were found in the intervention group as participants were in a position to describe the beneficial psychological effects the virtual window provided to them. This study is an example of how embedded mixed methods designs can be used to identify not only if interventions work, but crucially why they might work.

Explanatory design

The explanatory design is a sequential mixed methods design usually consisting of a larger quantitative phase followed by a smaller qualitative phase, the aim of which is to follow up and explain the quantitative findings. Data are collected and analyzed separately and then integrated in the interpretation phase of the study. The distinguishing attribute of a sequential explanatory mixed methods design is that the second phase of the study arises as a result of, or in response to, the findings of the first phase (Creswell et al., 2003). This design is particularly useful for following up on quantitative results that were either unexpected or that require more detailed explanation by sub-groups of the original sample. The main drawback to this mixed methods design is that it may take considerable time and resources to undertake distinct phases of a study (Ivankova, Creswell, & Stick, 2006) and therefore may be more difficult for individual researchers to complete. Petros (2012) utilized a sequential explanatory design to investigate the support needs of older caregivers to family members affected by HIV and AIDS in South Africa. In this study, a survey was first used to assess the extent of the problem of caregiving by older persons. A qualitative phase then provided the opportunity to gather data to better understand the burden of caregiving on these caregivers and areas in which they needed formal support. The qualitative phase allowed for responses to be probed and for insight to be gained thereby increasing comprehensiveness of the results and enhancing understanding of HIV/AIDS-related caregiving.

Exploratory design

The exploratory design is also a two-part design beginning with the qualitative phase which is typically employed to develop or inform the quantitative methods (Creswell & Plano Clark, 2007). This design is best employed when little is known about the variables to be studied. The

qualitative phase is commonly used to develop or try out an instrument in a specific group, to determine important variables to be studied or to develop hypotheses for testing based on emerging theory. A sequential exploratory mixed methods design was utilized by O'Leary & Ni Mhaolrunaigh (2011) to study the information-seeking behaviors that nurses use to inform decision making. Exploratory qualitative interviews of 29 nurses were employed in the first phase which yielded data contributing to the development of a questionnaire subsequently distributed to 1,356 nurses.

More recent work has advanced this basic typology by considering the degree of integration and the number of strands included in the design (e.g., Teddlie & Tashakkori, 2009). Regardless of the specific typology guiding researchers, it is clear that mixed methods designs in all of their forms have much to offer the quantitative researcher.

Challenges of mixed methods research

Mixed methods research is still relatively new. Continuing vigorous dialogue about foundations and methods focuses on challenges including data analysis, quality assessment, philosophical foundations, logistics, and tokenism.

Data analysis

Connection and integration of quantitative and qualitative data occurs during the data analysis phase (Teddlie & Tashakkori, 2009). Choice of analysis approach is guided by the rationale for mixing quantitative and qualitative research and the mixed method design utilized. Analysis of qualitative and quantitative data may be done separately, or may be synthesized during the analysis phase.

To date, many mixed methods studies in nursing and the health sciences have analyzed quantitative and qualitative data separately with integration only occurring at the point of conclusion and discussion. In some designs (e.g., most sequential designs), qualitative and quantitative data analysis occurs separately because detailed design of the second phase of the project is guided by results of the analysis of data from the first phase. In such cases where separateness of the quantitative analyses and qualitative analyses is crucial, a common understanding is still achieved through the integration of conclusions from quantitative findings and qualitative findings into overarching meta-inferences (Onwuegbuzie & Combs, 2010; Teddlie & Tashakkori, 2009). However, as Bazeley (2009) noted, while this separateness of analysis might sometimes be appropriate to the purpose of the study, the lack of integration results in lost opportunity for a richer, more depth analysis of the data. Advances in mixed methods data analysis techniques see a focus on synthesis of data. Rather than analyzing separate qualitative and quantitative datasets, numerical data may be 'qualitized' or textual data may be 'quantitized' (Bazeley, 2009; Teddlie & Tashakkori, 2011). In addition, qualitative data from individual cases can be linked to demographic data or responses on a survey (Bazeley, 2012). A number of qualitative analysis software programs such as NVivo and MAXQDA facilitate this. While this form of analysis will not be suitable for all mixed methods studies, it extends integration from planning and interpretation to include the analysis phase of a study. Issues regarding data analysis in mixed methods studies have received growing attention in recent years (Teddlie & Tashakkori, 2010) and it is likely that this trend will continue as the design develops further. Additional reading on the complex issue of data analysis and integration in mixed methods research can be found in Teddlie and Tashakkori (2009), Onwuegbuzie and Combs (2010), and Bazeley (2009, 2010) amongst others.

Quality

Quality assessment is another challenge in mixed methods studies. A basic question is whether quality can be achieved by assuring that quantitative and qualitative parts separately utilize quality standards appropriate to that methodology, or whether mixed methods needs its own quality criteria (O’Cathain, 2010). In reality, elements of both are required to ensure a quality mixed methods study. Quality mixed methods studies require that the individual quantitative and qualitative parts of the study are conducted in a rigorous manner in accordance with the established principles of that methodology (Wisdom, Cavaleri, Onwuegbuzie, & Green, 2012). However, while high-quality quantitative and qualitative components are necessary, they are not sufficient for a mixed methods study (Tashakkori & Teddlie 2008) because quality of the whole is more than the sum of its quantitative and qualitative parts (Creswell & Plano Clark, 2007). Consequently, there is the need for a quality assessment of the whole mixed methods study and an increasing tendency is to focus on evaluating the quality of meta-inferences, which are integrated understandings derived from *both* the quantitative and qualitative components (Tashakkori & Teddlie, 2008). Meta-inferences go beyond what is learned from the quantitative and qualitative components in isolation and the quality standards of meta-inferences also go beyond those of each individual strand (Tashakkori & Teddlie, 2008). In brief, the evaluation of the quality of inferences is focused on the two central components of design quality (appropriateness of the procedures) and interpretative rigor (authenticity of conclusions from a study) (O’Cathain, 2010; Tashakkori & Teddlie, 2008).

Quality assessment in mixed methods studies however is often hindered by a lack of transparency in the reporting of these studies (O’Cathain, Murphy, & Nicholl, 2008), with few mixed methods authors justifying the use of a mixed methods design or clearly describing the components of a study unique to mixed methods (Wisdom et al., 2012). This raises the question of what essential elements should be reported in a mixed methods study. Clearly, research reports should identify the methodological processes relevant to both the quantitative and qualitative components. In addition, issues inherent to a mixed methods study need also be reported. Guidelines for reporting a mixed methods study (e.g., Good Reporting of a Mixed Methods Study [GRAMMS]; O’Cathain et al., 2008) suggest including:

- A justification for using a mixed methods design;
- A description of the design in terms of the purpose, priority, and sequence of methods;
- A description of the methods including sampling, data collection, and analysis;
- A description of where and how integration has occurred;
- Identification of any limitations of using a mixed methods approach;
- A description of any insights gained from mixing methods.

In addition, there is a requirement to appraise the quality of mixed studies reviews (MSRs) which concomitantly examine qualitative, quantitative, and mixed methods primary studies (Pluye, Gagnon, Griffiths, & Johnson-Lafleur, 2009). Similar to the work of O’Cathain et al. (2008), Pluye et al. (2009) suggest a simple scoring system to appraise the quality of MSRs that includes an appraisal of the quantitative and qualitative studies and components, with reference to quality standards appropriate to those traditions, in addition to an appraisal of the mixed methods studies and components with specific reference to justification for the combination of methods and integration of findings from mixed methods studies.

Philosophical foundations

On a philosophical level, challenges still exist in justifying the merger of quantitative and qualitative research methods in the same study. Denzin (2012) cautioned that it is a mistake to forget about the core differences that exist between quantitative and qualitative methods. On a practical level, researchers have been fruitfully combining quantitative and qualitative methods for some time now (Teddle & Tashakkori, 2011), and as O’Cathain (2009) noted, mixed methods designs are justified on pragmatic rather than ideological grounds to ensure that complex health care research questions can be studied.

Logistics

Logistical challenges related to cost, expertise, and time present themselves with the use of mixed methods designs. Mixed methods studies may be more expensive requiring as they do the collection and analysis of two data sets. Solo researchers conducting a quality mixed methods study must have competency not only in the use of both quantitative and qualitative methods but also in the third methodology of mixed methods. For this reason, mixed methods research may be particularly suited (although not limited to) teams of researchers. Finally, the time required to conduct a mixed methods with sometimes two distinct phases is normally longer than conducting mono-method studies. This is a consideration for doctoral students or those working to tight externally set timelines (Teddle & Tashakkori, 2011).

Tokenism

A final challenge is the sometimes tokenistic way mixed methods research is utilized, in which (usually) a qualitative method is half-heartedly included in a quantitative study often to improve the marketability of a project (Bergman, 2008). Such utilization of mixed methods designs can stymie the true potential of qualitative research to fruitfully contribute to the outcomes of a predominately quantitative study.

Conclusion

The increased use of mixed methods designs by nurse researchers is to be welcomed for its potential to increase the applicability of research findings to the real-world of patients and carers. It is becoming clear that mixed methods research works at the level of practice, producing findings that are person centered *as well as* variable centered. In this light, there is an acceptance that mixed methods research provides an appropriate methodology to address the complex health problems frequently faced by the nursing discipline.

References

- Andrew, S., & Halcomb, E. J. (2009). Future challenges for mixed methods research in nursing and the health sciences. In S. Andrew & E. Halcomb (Eds.), *Mixed methods research for nursing and the health sciences* (pp. 217–224). Chichester: Wiley-Blackwell.
- Bamberger, M., Rao, V., & Woolcock, M. (2010). Using mixed methods in monitoring and evaluation: Experiences from international development. In A. Tashakkori & C. Teddle (Eds.), *Handbook of mixed methods in social and behavioural research* (2nd ed.,) (pp. 613–641). Thousand Oaks, CA: Sage.

- Bazeley, P. (2009). Analysing mixed methods data. In S. Andrew & E. Halcomb (Eds.), *Mixed methods research for nursing and the health sciences* (pp. 84–118). Chichester: Wiley-Blackwell.
- Bazeley, P. (2010). Computer-assisted integration of mixed methods data sources and analyses. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research* (2nd ed.) (pp. 431–467). Thousand Oaks, CA: Sage.
- Bazeley, P. (2012). Integrative analysis strategies for mixed data sources. *American Behavioral Scientist*, 56, 814–828.
- Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71, 287–311.
- Beck, C. T. (Ed.). (2013). *Routledge international handbook of qualitative nursing research*. Oxford, UK: Routledge/Taylor & Francis.
- Bergman, M. M. (Ed.). (2008). Introduction: Whither mixed methods? In *Advances in mixed methods research: Theories and applications* (pp. 1–7). London: Sage.
- Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research*, 6, 97–113.
- Campbell, N. C., Murray, E., Darbyshire, J., Emery, J., Farmer, A., Griffiths, F., . . . Kinmonth, A. L. (2007). Designing and evaluating complex intervention to improve health care. *BMJ*, 333(7591), 455–459.
- Creswell, J. W. (2011). Controversies in mixed methods research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE handbook of qualitative research* (pp. 269–283). Thousand Oaks, CA: Sage.
- Creswell, J. W., Fetters, M. D., Plano Clark, V. L., & Morales, A. (2009). Mixed methods intervention trials. In S. Andrew & E. Halcomb (Eds.), *Mixed methods research for nursing and the health sciences* (pp. 161–180). Chichester: Wiley-Blackwell.
- Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.
- Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research design. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research* (pp. 209–240). Thousand Oaks, CA: Sage.
- Denzin, N. K. (2012). Triangulation 2.0. *Journal of Mixed Methods Research*, 6, 80–88.
- Denzin, N. K., & Lincoln, Y. S. (2005). Introduction: The discipline and practice of qualitative research. In *The SAGE handbook of qualitative research* (3rd ed.) (pp. 1–32). Thousand Oaks, CA: Sage.
- DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. on behalf of the PROMIS Cooperative Group. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care*, 45(5 Suppl. 1), S12–S21.
- Doyle, L., Brady, A.-M., & Byrne, G. (2009). An overview of mixed methods research. *Journal of Research in Nursing*, 14, 175–185.
- Feilzer, M. Y. (2010). Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. *Journal of Mixed Methods Research*, 4, 6–16.
- Flemming, K. (2007). The knowledge base for evidence-based nursing: A role for mixed methods research. *Advances in Nursing Science*, 30, 41–51.
- Giddings, L. S. (2006). Mixed-methods research: Positivism dressed in drag? *Journal of Research in Nursing*, 11, 195–203.
- Gorad, S. (2010). Research design, as independent of methods. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research*. (2nd ed.) (pp. 237–251). Thousand Oaks, CA: Sage.
- Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.
- Greene, J. C., & Caracelli, V. J. (1997). Defining and describing the paradigm issue in mixed-method evaluation. *New Directions for Program Evaluation*, 4, 5–17. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/ev.1068/pdf>
- Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11, 255–274.
- Henly, S. J. (2007). Lost in time. The person in nursing research [Editorial]. *Nursing Research*, 56, 147.
- Howe, K. R. (1988). Against the quantitative–qualitative incompatibility thesis or dogmas die hard. *Educational Researcher*, 17, 10–16.
- Ivankova, N. V., Creswell, J. W., & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field Methods*, 18, 3–20.

- Johnson, B., & Gray, R. (2010). A history of philosophical and theoretical issues for mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research* (2nd ed.,) (pp. 69–94). Thousand Oaks, CA: Sage.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14–26.
- Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Towards a definition of mixed methods research. *Journal of Mixed Methods Research*, 1, 112–133.
- Leech, N. L., & Onwuegbuzie, A. J. (2009). A typology of mixed methods designs. *Quality and Quantity*, 43, 265–275.
- Marcus, S. M., Stuart, E. A., Wang, P., Shadish, W. R., & Steiner, P. M. (2012). Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. *Psychological Methods*, 17, 244–254.
- McCabe, C., Roche, D., Hegarty, E., & McCann, S. (2011). “Open Window”: A randomized trial of the effect of new media art using a virtual window on quality of life in patients’ experiencing stem cell transplantation. *Psycho-Oncology*. doi:10.1002/PON.2093
- Morgan, D. L. (2007). Paradigms lost and pragmatism regained. Methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research*, 1, 48–76.
- Muncey, T. (2009). Does mixed methods constitute a change in paradigm? In S. Andrew & E. Halcomb (Eds.), *Mixed methods research for nursing and the health sciences* (pp. 13–30). Chichester: Wiley-Blackwell.
- O’Cathain, A. (2009). Mixed methods research in the health sciences. *Journal of Mixed Methods Research*, 3, 3–6.
- O’Cathain, A. (2010). Assessing the quality of mixed methods research: Toward a comprehensive framework. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research* (2nd ed.,) (pp. 531–555). Thousand Oaks, CA: Sage.
- O’Cathain, A., Murphy, E., & Nicholl, J. (2008). The quality of mixed methods studies in health services research. *Journal of Health Services Research Policy*, 13, 92–98.
- O’Leary, D. F., & Ni Mhaolrunaigh, S. (2011). Information-seeking behaviour of nurses: Where is information sought and what processes are followed? *Journal of Advanced Nursing*, 68, 379–390.
- Onwuegbuzie, A. J. (2002). Why can’t we all get along? Towards a framework for unifying research paradigms. *Education*, 122, 518–530.
- Onwuegbuzie, A. J., & Combs, J. P. (2010) Emergent data analysis techniques in mixed methods research: A synthesis. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research* (2nd ed.,) (pp. 397–430). Thousand Oaks, CA: Sage.
- Ossebaard, H. C., Seydel, E. R., & van Gemert-Pijnen, L. (2012). Online usability and patients with long-term conditions: A mixed-methods approach. *International Journal of Medical Informatics*, 81, 374–387.
- Parahoo, K. (2006). *Nursing research: Principles, process and issues* (2nd ed.). Houndmills: Palgrave Macmillan.
- Petros, S. G. (2012). Use of a mixed methods approach to investigate the support needs of older caregivers to family members affected by HIV and AIDS in South Africa. *Journal of Mixed Methods Research*, 6, 275–293.
- Pluye, P., Gagnon, M-P, Griffiths, E., & Johnson-Lafleur, J. (2009). A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in Mixed Studies Reviews. *International Journal of Nursing Studies*, 46, 529–546.
- Polit, D. F., & Beck, C. T. (2006). *Essentials of nursing research: Methods, appraisals, and utilization* (6th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.
- Rallis, S. F., & Rossman, G. B. (2003). Mixed methods in evaluation contexts: A pragmatic framework. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research* (pp. 491–512). Thousand Oaks, CA: Sage.
- Relton, C., Torgerson, D., O’Cathain, A., & Nicholl, J. (2010). Rethinking pragmatic randomised controlled trials: Introducing the “cohort multiple randomised controlled trial” design. *BMJ*, 340, c1066.
- Sandelowski, M. (1996). Using qualitative methods in intervention studies. *Research in Nursing and Health*, 19, 359–364.
- Schatz, E. (2012). Rationale and procedures for nesting semi-structured interviews in surveys or censuses. *Population Studies*, 66, 183–195.
- Song, M.-K., Sandelowski, M., & Happ, M. B. (2010). Current practices and emerging trends in conducting mixed methods intervention studies in the health sciences. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research* (2nd ed.,) (pp. 725–747). Thousand Oaks, CA: Sage.

- Streubert Speziale, H. J., & Carpenter, D. R. (2007). *Qualitative research in nursing: Advancing the humanistic imperative* (4th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.
- Tashakkori, A., & Creswell, J. W. (2007). The new era of mixed methods. *Journal of Mixed Methods Research*, 1, 3–7.
- Tashakkori, A., & Teddlie, C. (2008). Quality of inferences in mixed methods research: Calling for an integrative framework. In M. M. Bergman (Ed.), *Advances in mixed methods research: Theories and applications* (pp. 101–119). London: Sage.
- Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools*, 13(1), 12–28.
- Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioural sciences*. Thousand Oaks, CA: Sage.
- Teddlie, C., & Tashakkori, A. (2010). Overview of contemporary issues in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research* (2nd ed.,) (pp. 1–41). Thousand Oaks, CA: Sage.
- Teddlie, C., & Tashakkori, A. (2011). Mixed methods research: Contemporary issues in an emerging field. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE handbook of qualitative research* (pp. 285–299). Thousand Oaks, CA: Sage.
- Twinn, S. (2003). Status of mixed methods research in nursing. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research* (pp. 541–556). Thousand Oaks, CA: Sage.
- Vaughan Dickson, V., Lee, C. S., & Riegel, B. (2011). How do cognitive function and knowledge affect heart failure self-care? *Journal of Mixed Methods Research*, 5, 167–189.
- West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., . . . Mullen, P. D. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, 98, 1359–1366.
- West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*, 15, 18–37.
- Wisdom, J. P., Cavaleri, M. A., Onwuegbuzie, A. J., & Green, C. A. (2012). Methodological reporting in qualitative, quantitative, and mixed methods health services research articles. *Health Services Research*, 47, 721–745.

GLOBAL GENERATION AND DISSEMINATION OF NURSING SCIENCE

Molly C. Dougherty and Nalini Jairath

Nursing science, exemplified through nursing research, evidence-based practice (EBP), and translational science initiatives, is bounded universally by the meta-constructs of health, nursing, person, environment (Flaskerud & Halloran, 1980), and time (Henly, Wyman, & Findorff, 2011). Global dissemination of nursing science involves sharing conceptualizations and paradigms, methodologies and measurement approaches, and data-based findings through varied modes of communication. This chapter provides an overview and analysis of current status, practices, challenges, and trends in global generation and dissemination.

Dissemination

Current status of global dissemination

The protean nature of knowledge development is best understood in light of the 21st-century global context for nursing as well as historical antecedents in nursing and science. Dating to the 18th century, scientific thought – most notably in the United Kingdom (UK), continental Europe, and the United States – flourished as a result of growth and funding of universities, national science policies, and institutionalization of the scientific method. These factors gave rise to the industrial revolution, colonialism, and, later, English as a shared language of science and the digital age.

Quantification of global dissemination of nursing science may be addressed through bibliometric analysis of information transfer pathways from original sources to general, clinical, and functional areas of nursing. Journal publications relative to other original sources have the highest transfer rate. In the nursing literature, citation analysis studies have indicated that 65% of total sources cited come from journals, 24% from books, and 4% from government sources (Allen, Jacobs, & Levy, 2006).

The accuracy of bibliometric analysis depends upon the adequacy of search strategies, specification of appropriate databases for nursing content areas, and identification of intersection points between nursing and other disciplines. The CINAHL database provides the best coverage for nursing journals. PubMed/MEDLINE provides average coverage for nursing and ranks with Embase and the Science Citation Index for biomedical journals, and PsycINFO and the Social Sciences Citation Index are needed for complete coverage of psychosocial topics (Allen et al.,

2006). Intersection points may occur when researchers hold multiple disciplinary affiliations or participate in trans-disciplinary initiatives (Jairath, 2007). Intersections may occur also when the documented high citation rate for literature from other disciplines (Allen et al., 2006) results in failure to use and cite relevant nursing literature. In the maternal-child nursing literature, for example, citation analysis has shown that medical journals predominate (Oermann, Blair, Kowalewski, Wilmes, & Nordstrom, 2007).

Currently, English is the dominant language of science journals and databases (Lu, 2011; Vasconcelos, Sorenson, & Leta, 2007) and, in non-English-speaking countries, publication in high-ranking journals, research funding, and English proficiency are related (Man, Weinkauff, Tsang, & Sin, 2004). Highly reputable, easily accessible databases such as PubMed/MEDLINE, Google Scholar, Scopus, and Thomson Reuters disproportionately represent Western origin and English language journals (Hofman, Kanyengo, Rapp, & Kotzin, 2009), although efforts are being made to include other regional literature in databases (Testa, 2011). These databases provide raw data for bibliometrics and further apparent linkages between quality and English language publication. Journal impact factor uses only Thomson Reuters databases (Freda, 2006), and citation analysis heavily emphasizes Thomson Reuters Web of Science, Google Scholar, and Scopus databases.

The perceived differences in quality between English and non-English language publications may reflect variations in scholarly productivity and infrastructure support. The growth rate of publications in non-English journals is generally lower (Loria & Arroyo, 2005). Regional and income differences in publication patterns exist. In a comprehensive study of the Latin American nursing research literature (Mendoza-Parra, Paravic-Klijn, Munoz-Munoz, Barriga, & Jimenez-Contreras, 2009), only 2% of researchers form the most productive group of authors. Further, despite rapid annual new journal growth, low journal longevity adversely affects consideration for inclusion in catalogs, indexes, or bibliographic databases. In Africa, while scholarship increased, single authorships predominated and few clinical studies were published (Adejumo & Lekalakala-Mokgele, 2009). In general, low- and middle-income countries have lower manuscript submission and publication rates, and low-income countries exhibit extremely low rates of publication in indexed journals (Mari et al., 2010).

Initiatives to address these issues are underway. The World Health Organization (WHO) has explored innovative mechanisms for improving quality and dissemination of national medical publications. A partnership program pairing the editors of four sub-Saharan African journals with UK/U.S.-based journals has been funded by the Fogarty International Center (Hofman et al., 2009). Global publishing groups (e.g., Elsevier, Springer, Blackwell, Nature) are collaborating to advance the international profile of Chinese English-language journals and facilitate publication by Chinese scholars in joint journals (Lu, 2011). Traditional publishers are using creative approaches to allow greater online access to archival material. Additionally, researchers are using multiple options for dissemination, including publishing evidence-based summaries, open access journals such as those of the nonprofit Public Library of Science (www.plos.org/) and the BioMed Central journal family, and fast-track publication mechanisms such as preprint online publications.

Dissemination and nursing education

Nursing education is recognized as an important precondition for the development of a critical mass of scientists, educators, and clinicians prepared to address theoretically grounded problems, apply scientific methods, and use evidence to change practice. Currently, inclusion of theory, research, and EBP in the curricula of all programs preparing entry-level nurses is advocated

nationally (i.e., National Council of State Boards of Nursing, 2010) and internationally (WHO, 2009), and additional content is advocated for advanced practice roles in the U.S. (American Association of Colleges of Nursing, 2011).

At the doctoral level, education advancing nursing science transcends national borders. From the initial U.S. programs of the 1950s (Murphy, 1985), doctoral education spread in the 1970s to England and Ireland (McKenna & Cutcliffe, 2001) and now spans five continents and at least 32 nations (International Network for Doctoral Education in Nursing, n.d.). Currently, three models of doctoral education exist: (a) a Doctor of Philosophy (PhD) model featuring extensive coursework and some type of comprehensive or qualifying exam, followed by dissertation research (the U.S. and Asian model); (b) a PhD model featuring limited coursework but more extensive dissertation research, mentorship, and the opportunity to conduct extensive research (the UK/European Model); and (c) a Doctor of Nursing Practice model (Edwardson, 2010) featuring a highly prescribed curriculum and a mandatory EBP project. Exponential growth of PhD programs has raised quality concerns related to faculty shortages, access to adequate research funding (Ketefian, Davidson, Daly, Chang, & Srisuphan, 2005; Minnick, Norman, Donaghey, Fisher, & McKirgan, 2010), and program responsiveness to societal needs. One concern is that close student–mentor research alignment may diminish scientific advancement in new domains (Kehm, 2006). Furthermore, existing models may not equip nurses adequately with the statistical expertise to use recent advances in applied statistics (Hayat, 2012), health services research (Jones & Lusk, 2002), or health trajectory research (Henly et al., 2011).

Dissemination and advancing practice

Global dissemination has the potential to advance EBP rapidly by increasing the availability of pertinent research and maximizing the strengths of research conducted in different countries. Polit and Beck (2009) identified clinically meaningful and statistically significant differences in study characteristics, study participants, and authorship by world region. Patient-focused studies were more common in the U.S. and Asian countries, and nurse-focused studies were more prevalent in Europe, Australia, and Canada. In the UK, Norway, and Sweden, qualitative methods were used more frequently than in Asia, where quantitative methods predominated.

Evidence-based practice benefits from testing and evaluation of evidence using interpretative, evaluative, and hypothesis- or theory-testing research. Thus, quantitative research constitutes a higher level of evidence than qualitative research and randomized clinical trials (RCT) are the pinnacle of evidence pyramids. Systematic reviews, meta-synthesis, and meta-analysis methodologies have been identified also as appropriate evidence sources when RCT data is unavailable or insufficient (Mantzoukas, 2009). Descriptive studies, when used to generate EBP recommendations, have significant limitations, including making non-data-based inferences and associations and weak global applicability.

When examined globally, the overall state of the science suggests a low rate of advancement in EBP. Mantzoukas (2009), using 2,574 articles from 10 highly ranked nursing journals, showed that in the years 2000 through 2006, RCTs were rare: only 4% of studies reviewed used experimental methodologies with control and randomization and 37% implemented qualitative methods. Beckstead (2009) used Mantzoukas's data to analyze the correlation between the methodology used (quantitative vs. qualitative) and journal impact factor (measured using citation analysis). Statistically significant, weak positive correlations ($r = .23$) established a link between quantitative research and the published literature, but statistically significant negative correlations ($r = -.29$) indicated that qualitative research articles were not linked. Based on the principle that

science builds on prior published research, Beckstead (2009) concluded that scientific research in nursing is progressing unevenly as a function of the category of issues being studied, types of methods employed, and evidence being generated and published. In addition, because qualitative research arguably represents a beginning, as compared to a fully developed research program, the immediate contribution to EBP is lesser.

Benefits of a global perspective on nursing

Nursing practice environments and workforce issues

Nursing health service research is used to address the practice environment. Some research addresses inherently international issues such as global nurse migration. Other research, such as studies on the nursing practice environment (NPE), addresses nurse and patient satisfaction or national and regional issues that shape nursing science globally.

Studies on nurse recruitment and retention drew attention to global nurse workforce shortages and nurse migration, ultimately leading to efforts to improve the NPE. The International Council of Nurses (ICN) and the Commission on Graduates of Foreign Nursing Schools jointly sponsor the global resource center on nurse migration and a repository of relevant research available through their website (www.intlnursemigration.org/sections/research/publications.shtml). A framework to support the integration of international nurses into host countries (Adams & Kennedy, 2006) has been developed. Policy papers have addressed return migration of nurses (Haour-Knipe & Davies, 2008) and nursing self-sufficiency in the global context (Little & Buchan, 2007).

Studies on the NPE have contributed to the understanding of nursing sensitive measures, patient outcomes, and various features of the hospital work setting. Seminal work was focused on characteristics of U.S. Magnet hospitals, hospitals that meet specific criteria for institutional excellence, high-quality nursing care, a research-EBP culture, and known success in attracting and retaining nurses (Kramer & Schmalenberg, 1988; McClure, Poulin, Sovie, & Wandelt, 1983).

The International Hospital Outcomes Study, in which well-grounded conceptualization, measures, and instruments were used, represented another major advance in global research on patient and nurse outcomes. The study included interdisciplinary teams from select areas of North America, the UK, Germany, and New Zealand (Clarke & Aiken, 2008). Despite identifying similarities across aspects of hospital nurses' work experiences, remarkable differences were identified within countries in working conditions and resources needed for high-quality practice. Consistent evidence existed to show connections between modifiable features of NPE, patient well-being, and factors influencing workforce stability (Clarke & Aiken, 2008). These studies and others on patient safety contributed to a paradigm for the study of various features of patient outcomes and NPE globally.

The paradigm incorporates patient outcomes, nurse satisfaction, NPE, and hospital characteristics; fits hospital nursing settings across the world; and has influenced nursing research strongly in the broad area of NPE. It has been used in studies from many countries, including countries where English is a second language and low- and middle-income countries. Research has included an 11-country survey of the expectations and needs of nurses in the workplace (DeCola & Riggins, 2010) and a seven-country study of differences regarding control over practice between northern (Europe and U.S.) and Mediterranean countries (Papastavrou et al., 2012). Global concerns around workload issues, cost-saving strategies, and restructuring have fostered a body of research consistent with the NPE paradigm on rationing of nursing care (e.g., extent to which interventions and care are withheld; Schubert, Glass, Clarke, Schaffert-Witvliet, & De

Geest, 2007). The influence of the ICN's efforts to identify positive NPE and stimulate research internationally is evident in studies of NPE from Taiwan (Lee, Pai, & Yen, 2008) and Spain (Garcia-Garcia, Ramos, Serrano, Ramos Cobos, & Souza, 2011). Together, the studies demonstrate the paradigm applicability to regional settings and the seemingly universal practice environment concerns of nursing.

Importantly, NPE and patient outcomes research has drawn the attention of policy-making bodies and funding agencies. The NPE and patient outcome research have placed concerns of nursing on national healthcare agendas, indicating that, globally, nursing research is building on nursing science. Issues related to national healthcare policy and the cost to nations is near the top of national agendas. This body of research shows that economy in healthcare is possible with attention to nurse-sensitive patient outcomes and NPE. Research in this and other areas grounded in solid foundational studies is likely to bring resources needed to advance nursing science globally.

Global dissemination of clinical research

Clinical research that has global impact requires nursing interventions to transcend culture and country. One such intervention is *Kangaroo* mother care (KMC), which is based on skin-to-skin contact between mother and infant, exclusive breastfeeding, and support to the mother–infant pair (www.skintoskincontact.com/an-introduction.aspx). Repeated RCTs have demonstrated the value of KMC in diverse care settings for optimizing newborn (including preterm infants) physiologic health and bonding, and lowering costs.

Researchers in KMC have maximized global dissemination by targeting researchers, clinicians, and consumers. Approaches include the U.S.-based International Kangaroo Care Awareness Day, the U.S. Institute for Kangaroo Care (www.kangaroocareusa.org/About_Us.html), and an online Kangaroo Care Bibliography (www.skintoskincontact.com/susan-ludington.aspx) containing outcome charts with links to published studies. The online bibliography permits cross-referencing between outcome categories such as apnea and specific indicators such as cortisol and acquisition of content regarding pertinent KMC studies.

Ludington, a leading KMC researcher, and her research group have promoted the perpetuation of KMC research actively by providing international consultation services as well as by involving nurses from other countries in KMC as predoctoral, post-doctoral, and visiting scientists. Preparation in KMC research teams has helped these researchers develop region-specific research. The net result of these dissemination efforts has been worldwide adoption of KMC into healthcare policy and global initiatives such as UNICEF's Baby-Friendly Initiative.

Characteristics of nursing science with global impact

The examples of research on NPE and KMC point to the characteristics of nursing science that foster global reach and impact: The problem under study is widely present, even universal, and it has a major impact on health and healthcare costs. The research is built within a paradigm that can be adapted to many care settings around the world. The results of research highlight factors or interventions that make a difference in mortality, morbidity, and cost of healthcare. The science is strong, stands up under critical scrutiny, and is replicated sufficiently often to demonstrate generalizability. The research is disseminated widely and appeals to an interdisciplinary audience of scientists, policymakers, clinicians, and consumers. Most important, the influence on policy is sufficiently strong that the problem becomes a priority for funding, which permits nursing science in the area to advance.

Challenges to global dissemination

Global dissemination of nursing science involves the mutual interchange of scientific thought among countries and regions of the world. Major challenges that affect nursing science dissemination and dissemination of science generally are described in the following section.

National science policy and funding

Nations differ globally in the ways science policy and funding priorities evolve. Often, healthcare systems arising from highly varied wealth of nations result in priority areas differing so greatly that it is difficult to identify common ground. A common ground is essential; global nursing science development requires financial support, a research infrastructure, and authority from funding agencies with disparate agendas and priorities to pursue research.

Nursing science issues, distinct from nursing issues in general, must be seen as central to national healthcare to secure a place on national science agendas. Because most nursing science is supported nationally, globally relevant research is undervalued unless presented within the context of national healthcare agendas and shared paradigms. When policymakers accept the centrality of nursing issues and nursing science to national healthcare, the status of nursing increases and support for nursing science follows.

Other organizations that support research and set agendas about research priorities also influence the development of nursing science. Some are global (ICN and Sigma Theta Tau International) and have a regional presence. Others, such as the International Collaboration for Community Health Nursing Research and the Nursing Research Institute, WHO Europe Collaborating Centre for Nursing, are regionally based but address global nursing concerns. Still others explicitly address more narrow national concerns, but play a prominent role in advancing nursing science policy because of the urgency of the problems addressed as well as highly coordinated dissemination efforts to facilitate more rapid translation of research into practice. These organizations include military organizations such as the U.S. Tri-Service Nursing Research Program, which solely funds nursing research related to the U.S. military. But, this research often has broader application and is available through open-access initiatives.

Faith-based organizations, many of which are globally based, have major interests in the direction and content of science. They may advance nursing science through their support for low-cost nursing educational programs, including university programs and nursing faculty scholarship, and translational science within faith-based national healthcare networks (Curry, 2010; Wall, 2001). In nations where the separation between church and state is unclear, the influence of religion on research agendas may be more direct and may affect knowledge development adversely in controversial areas.

Linguistic challenges in publication

English language dominance facilitates development of global linkages around clinical concerns and building global relationships and initiatives. Conversely, it imposes an additional burden on non-English-speaking scientists who seek broader dissemination of their research in highly ranked journals, and may marginalize nurses, most notably clinical nurses who are unable to access or understand English publications.

English language dominance has contributed to an overwhelming volume of scientific information, which is often fragmented and of variable quality. Professional and specialty organizations, university libraries, and governmental groups vet the literature, provide guides to differing

levels of knowledge, and develop multifaceted approaches such as that used for KMC to foster dissemination in English and national language literature.

Culture plays an important role in beliefs and attitudes about health, attribution of causal relationships, and the posited mechanisms for action of nursing interventions. Though necessary for scientific advancement, synchronizing different health belief systems and perspectives on nursing issues requires intense involvement and may appear to invalidate cultural beliefs. However, as exemplified by research regarding the *good nurse* construct, common ground can be found. This construct has been addressed in U.S. (Catlett & Lovan, 2011; Smith & Godfrey, 2002), UK (Brady, 2009), and Irish (Fealy, 2004) research. Additionally, research in Hong Kong, mainland China, Korea, and Japan (Izumi, Konishi, Yahiro, & Kodama, 2006; Pang, Yahairo, & Chan, 2009) has contributed to the understanding of the good nurse construct within the greater Southeast Asia culture.

There is broad-based concern among health experts about the disparity between generation of knowledge and disease burden in low- and middle-income countries compared to high-income countries. The challenge for nursing is that nursing science in most of the world is immature and issues for nursing in low- and high-income countries are widely disparate. Providing nursing science expertise to low-income countries is complicated by the relatively high ratio of nurses to consumers, especially nurses with advanced preparation in science, in low- and middle-income countries globally and by the relatively low investment in nursing science throughout the world. Clearly, it will be necessary to improve dissemination of nursing science and added attention to nursing science in national languages is needed.

Strategies to improve the availability of science dissemination in national languages include journal editors and publishers sponsoring reproduction of previously presented research reports in clinical journals, and increased use of companion websites and resources to increase access to research. Release of copyrights by research journals (to promote dissemination of science and to allow dual publication of research in English and in national languages) is a strategy that has worked in some disciplines. The open access movement promoted by health science center librarians, scientists in many disciplines, and the National Institutes of Health has improved the global dissemination of nursing science.

Though underutilized currently by scientists, social media such as Twitter, blogs, Facebook, and other emerging modalities are likely to influence global dissemination in the future. Social media may be used to generate interest in a scientific topic and to direct followers with shared interests to scholarly sites for complete information. Scientists who use these forms of communication will need to determine the potential benefits in light of the additional time required to engage these venues to reach a focused audience with similar interests.

The limited number of nurse scientists

The number of doctorally prepared nurses globally is distributed unevenly throughout the globe. Many of these nurses are in administrative or clinical positions in which they lack the resources to conduct substantive research. Therefore, the challenge is to increase the cadre of nurses with advanced degrees, especially in low- and middle-income countries where the need is great. Globally, advancing education for nurses, fostering career paths that include preparation in science, and placing nurses in settings where nursing science is supported are essential.

Academic institutions, professional organizations, and funding agencies are able to foster research cultures that support the development of individual nurse researchers and research programs. Creation of academic-clinical partnerships and support for research teams comprised of nurses with various educational backgrounds and levels of proficiency is helpful. Effective strategies must fit national settings and embrace a global perspective.

Dissemination, implementation science, and informatics

Science dissemination has been revolutionized by the World Wide Web, which has made communication nearly instantaneous globally. Still, the time lag from discovery to diffusion is long and the explosion of information increases the difficulty that scientists, nurses, and consumers have in using it. Strategies that combine the speed and access of the Web, and communication with specific highly interested audiences globally with knowledge about research processes and criteria for research quality are needed to maximize the dissemination of nursing science.

Regional nursing science partnerships are valuable because common cultural and environmental conditions give rise to major regional health problems, but regional linkages are impeded by a number of factors discussed earlier in the chapter. Strategies to advance nursing science regionally include international research collaborations to which individual programs of research may contribute and, when combined, can advance nursing science at the regional level.

Strategies to foster regional partnerships include the use of guidelines developed by international nursing organizations. Sigma Theta Tau International, WHO, and other organizations maintain websites that may be helpful in nurturing these partnerships. The dissemination of nursing science advances across multiple fronts with all approaches important to improving the yield of nursing science globally.

Conclusion

Several issues and challenges pertain to the global dissemination of nursing science. These include variations in the quality and utility of the underlying science, global workforce issues, and national science policy that does not provide adequately for research generation and dissemination. However, several notable collaborative exemplars demonstrate benefits of global dissemination and use of globally generated databases to improve practice.

References

- Adams, E., & Kennedy, A. (2006). Positive practice environments: Key considerations for the development of a framework to support the integration of international nurses. Retrieved from www.intlnursemigration.org/assets/pdfs/ICNM_Pos_Practice_Env.pdf
- Adejumo, O., & Lekalakala-Mokgele, L. (2009). A 2-decade appraisal of African nursing scholarship: 1986–2006. *Journal of Nursing Scholarship*, 41, 64–69.
- Allen, M. P., Jacobs, S. K., & Levy, J. R. (2006). Mapping the literature of nursing: 1996–2000. *Journal of the Medical Library Association*, 94, 206–220.
- American Association of Colleges of Nursing. (2011). Essentials of master's education in nursing. Retrieved from www.aacn.nche.edu/education-resources/MastersEssentials11.pdf
- Beckstead, J. W. (2009). What we write and what we cite: A comment on Mantzoukas (2009). *International Journal of Nursing Studies*, 46, 1161–1163.
- Brady, M. (2009). Hospitalized children's views of the good nurse. *Nursing Ethics*, 16, 543–560.
- Catlett, S., & Lovan, S. R. (2011). Being a good nurse and doing the right thing: A replication study. *Nursing Ethics*, 18, 54–63.
- Clarke, S. P., & Aiken, L. H. (2008). An international hospital outcomes research agenda focused on nursing: Lessons from a decade of collaboration. *Journal of Clinical Nursing*, 17, 3317–3323.
- Curry, B. D. (2010). Nursing education and the Catholic tradition. *Health Progress*, 91, 24–28.
- DeCola, P. R., & Riggins, P. (2010). Nurses in the workplace: Expectations and needs. *International Nursing Review*, 57, 335–342.
- Edwardson, S. R. (2010). Doctor of philosophy and doctor of nursing practice as complementary degrees. *Journal of Professional Nursing*, 26, 137–140.
- Fealy, G. M. (2004). "The good nurse": Visions and values in images of the nurse. *Journal of Advanced Nursing*, 46, 649–656.

- Flaskerud, J. H., & Halloran, E. J. (1980). Areas of agreement in nursing theory development. *Advances in Nursing Science*, 3, 1–7.
- Freda, M. C. (2006). Don Quixote, David and me. *Nursing Outlook*, 54, 58–59.
- Garcia-Garcia, I., Ramos, V. B., Serrano, J. L., Ramos Cobos, M. C., & Souza, A. (2011). Nursing personnel's perceptions of the organizational climate in public and private hospitals in Spain. *International Nursing Review*, 58, 234–241.
- Haoir-Knipe, M., & Davies, A. (2008). Return migration of nurses. Retrieved from www.intlnursemigration.org/assets/pdfs/ReturnmigrationA4.pdf
- Hayat, M. J. (2012). Statistics in nursing research. *Nursing Research*, 61, 147–148.
- Henly, S. J., Wyman, J. F., & Findorff, M. J. (2011). Health and illness over time: The trajectory perspective in nursing science. *Nursing Research*, 60(3 Suppl.), S5–S14.
- Hofman, K. J., Kanyengo, C. W., Rapp, B. A., & Kotzin, S. (2009). Mapping the health research landscape in Sub-Saharan Africa: A study of trends in biomedical publications. *Journal of the Medical Library Association*, 97, 41–44.
- Izumi, S., Konishi, E., Yahiro, M., & Kodama, M. (2006). Japanese patients' descriptions of "the good nurse": Personal involvement and professionalism. *Advances in Nursing Science*, 29, E14–E26.
- Jairath, N. (2007). Global health: The role of nursing research. *Nursing Research*, 56, 367–368.
- Jones, C. B., & Lusk, S. L. (2002). Incorporating health services research into nursing doctoral programs. *Nursing Outlook*, 50, 225–231.
- Kehm, B. M. (2006). Doctoral education in Europe and North America: A comparative analysis. Retrieved from www.portlandpress.com/pp/books/online/fyos/083/0067/0830067.pdf
- Ketefian, S., Davidson, P., Daly, J., Chang, E., & Srisuphan, W. (2005). Issues and challenges in international doctoral education in nursing. *Nursing and Health Sciences*, 7. Retrieved from http://deepblue.lib.umich.edu/bitstream/2027.42/88184/1/ketefian-issues_challenges_international_doctoral.pdf
- Kramer, M., & Schmalenberg, C. (1988). Magnet hospitals: Part I. Institutions of excellence. *Journal of Nursing Administration*, 18, 13–24.
- Lee, S., Pai, H. C., & Yen, W. J. (2008). [A study of perceived practice environment among nurses in Taiwan.] *Hu Li Za Zhi*, 55, 30–40.
- Little, L., & Buchan, J. (2007). Nursing self sufficiency/sustainability in the global context. Retrieved from www.intlnursemigration.org/assets/pdfs/SelfSufficiency_US.pdf
- Loria, A., & Arroyo, P. (2005). Language and country preponderance trends in MEDLINE and its causes. *Journal of the Medical Library Association*, 93, 381–385.
- Lu, J.-P. (2011). Is editing the roadblock to the internationalization of Chinese academic journals? *Journal of Zhejiang University, Science B*, 12, 503–506.
- Man, J. P., Weinkauff, J. G., Tsang, M., & Sin, D. D. (2004). Why do some countries publish more than others? An international comparison of research funding, English proficiency and publication output in highly ranked general medical journals. *European Journal of Epidemiology*, 19, 811–817.
- Mantzoukas, S. (2009). The research evidence published in high impact nursing journals between 2000 and 2006: A quantitative content analysis. *International Journal of Nursing Studies*, 46, 479–489.
- Mari, J. J., Patel, V., Kielsing, C., Razzouk, D., Tyrer, P., & Herrman, H. (2010). The 5/95 gap in the indexation of psychiatric journals of low- and middle-income countries. *Acta Psychiatrica Scandinavica*, 121, 152–156.
- McClure, M., Poulin, M., Sovie, M. D., & Wandelt, M. A. (1983). *Magnet hospitals: Attraction and retention of professional nurses*. Kansas City, MO: American Nurses Association.
- McKenna, H., & Cutcliffe, J. (2001). Nursing doctoral education in the United Kingdom and Ireland. *Online Journal of Issues in Nursing*, 6. Retrieved from www.nursingworld.org/MainMenuCategories/ANAMarketplace/ANAPeriodicals/OJIN/TableofContents/Volume62001/No2May01/ArticlePreviousTopic/UKandIrelandDoctoralEducation.aspx
- Mendoza-Parra, S., Paravic-Klijn, T., Munoz-Munoz, A. M., Barriga, O. A., & Jimenez-Contreras, E. (2009). Visibility of Latin American nursing research (1959–2005). *Journal of Nursing Scholarship*, 41, 54–63.
- Minnick, A. F., Norman, L. D., Donaghey, B., Fisher, L. W., & McKirgan, I. M. (2010). Defining and describing capacity issues in US doctoral nursing research programs. *Nursing Outlook*, 58, 36–43.
- Murphy, J. F. (1985). Doctoral education of nurses: Historical development, programs, and graduates. *Annual Review of Nursing Research*, 3, 171–189.
- National Council of State Boards of Nursing. (2010). 2010 NCLEX-RN® detailed test plan: Item writer/ item reviewer/nurse educator version. Retrieved from www.ncsbn.org/2010_NCLEX_RN_Detailed_Test_Plan_Educator.pdf

- Oermann, M. H., Blair, D. A., Kowalewski, K., Wilmes, N. A., & Nordstrom, C. K. (2007). Citation analysis of the maternal/child nursing literature. *Pediatric Nursing*, 33, 387–391.
- Pang, S.M.C., Yahairo, M., & Chan, H.Y.L. (2009). Knowing the patient and being a good nurse. In R. C. Locsin & M. J. Purnell (Eds.), *A contemporary nursing process: The (un)bearable weight of knowing in nursing* (pp. 445–461). New York, NY: Springer.
- Papastavrou, E., Efsthathiou, G., Acaroglu, R., Da Luz, M. D., Berg, A., Idvall, E., . . . Suhonen, R. (2012). A seven country comparison of nurses' perceptions of their professional practice environment. *Journal of Nursing Management*, 20, 236–248.
- Polit, D. F., & Beck, C. T. (2009). International differences in nursing research, 2005–2006. *Journal of Nursing Scholarship*, 41, 44–53.
- Schubert, M., Glass, T. R., Clarke, S. P., Schaffert-Witvliet, B., & De Geest, S. (2007). Validation of the Basel Extent of Nursing Care instrument. *Nursing Research*, 56, 416–424.
- Smith, K. V., & Godfrey, N. S. (2002). Being a good nurse and doing the right thing: A qualitative study. *Nursing Ethics*, 9, 301–312.
- Testa, J. (2011). The globalization of web of science. Retrieved from http://thomsonreuters.com/products_services/science/free/essays/regional_content_expansion_wos/
- Vasconcelos, S. M., Sorenson, M. M., & Leta, J. (2007). Scientist-friendly policies for non-native English-speaking authors: Timely and welcome. *Brazilian Journal of Medical and Biological Research*, 40, 743–747.
- Wall, B. M. (2001). “Definite lines of influence”: Catholic sisters and nurse training schools, 1890–1920. *Nursing Research*, 50, 314–321.
- World Health Organization. (2009). Global standards for the initial education of professional nurses and midwives. Retrieved from www.who.int/hrh/nursing_midwifery/hrh_global_standards_education.pdf

INDEX

- actigraphy 91–2
- adherence 92, 195, 246
- algorithms 221; CAIM 301–2; clustering 303; DataSqueezer 307; feature selection 299; Kohonen’s self-organizing map 307; miDataSqueezer 307–13; molecular dynamics method 320; for motion correction 347; rule-based 306; Sammon’s projection 319
- AMOS 131–2
- assignment mechanism 19, 214, 215–16, 229, 232–3; regular 216, 234
- asymptotically consistent estimator 125
- asymptotically distribution free (ADF) estimator 126
- average causal effect (ACE) 243–4
- bag 304–5
- balance 19, 213–14, 217, 232
- basal metabolic rate 100
- Bayesian approaches 22, 70, 71, 77, 164, 219–20, 254, 259, 288, 342, 376
- Bayesian information criterion (BIC) 143
- behavioral observation: definition 81–2; examples 82–3; molar concepts & molecular behaviors 84–5; observational accuracy 93–5; observational context 86–9; recording methods 89–93; temporal characteristics 86; time sampling 87–9
- behavioral observation topography of behavior 85
- bias: and blinding 205–6; comparative effectiveness studies 365; estimator 125–6, 144, 158, 182, 183, 218, 225, 229–30, 234, 242, 247–8, 253, 348, 352–3; observer 94; outcome reporting 367
- bias, selection: and confounders 241; in IRT 73; in joint models 254; in quasi-experiments 229–30; in safety research 390
- bibliometrics 423–4
- big data 19, 22, 286, 294
- biophysical measurement 92–112
- biospecimens 327–8, 333
- Bland-Altman plot 51–2
- blinding 205–6, 213
- blood flow measurement 110–12
- BOLD 107, 336, 343–5
- bootstrapping 314–15
- Braden scale 51, 156–66
- burden 47
- candidate gene study 328–9
- categorical observed variables in SEM 129
- causal effects 243–4, 361, 384, 388
- causal estimand 216–17
- causal inference: and moderator variables 241; in non-randomized studies 219, 225–7
- causal inference modes 217–21; Bayesian methods 219–20, 221; Fisherian mode 218; Neymanian mode 218–19
- causality 125, 213–21
- change 21, 38–9, 52–3, 139–40, 170–2, 206; minimally important change 52–3; reliable change 39; *see also* health trajectory
- change and variability models 170–84
- chemical energy 99–100
- classical test theory (CTT) 29–31, 58–9
- classifier 299–300; classifier errors 315–18
- clinimetrics 44–55
- Cochrane Collaboration 363–4
- coefficient alpha: computation 32–3; internal consistency 34–5; limitations 33–5
- comparative effectiveness research 3–4, 361
- complier average causal effect (CACE) 246
- concept in nursing science 17, 40
- conditional independence 69
- confidence interval 21, 38–9, 373
- confirmatory factor analysis (CFA) model 122–3

- confounder variable 238, 241–2, 244–5
 CONSORT 201, 205–7, 210
 construct validity 49
 content validity 49, 93
 controlled direct effect 243–4
 COSMIN 44
 cost-benefit analysis 398, 400–1
 cost consequence analysis 398, 400
 cost effectiveness analysis 398, 401
 cost minimization analysis 398, 400
 costs 405–7
 cost utility analysis 398, 402–3
 covariance(s) 19, 32–3, 120, 122, 123, 125, 140, 141
 covariates 238, 255; in general growth mixture models 141–2; in observational studies 217; in randomized experiments 213–14
 Cox proportional hazards model 255
 cross-validation 314

 data 16, 18, 268–9, 296, 330–1; analysis in mixed methods designs 417; coding 46–7, 59, 63, 84, 274–5, 283–4; elements 282; genotype 329 (*see also* biospecimens); mining 294; preprocessing 297; in safety research 387–8; secondary use 275, 286; in systematic review 367–8, 370; training and test 314
 data to wisdom model 267–8
 data warehouse 273–4
 decisions in clinical practice/policy 8–9, 376, 397, 401–2
 derivatives 172–3
 design: candidate gene association 328–9; difference in differences 234–5; experiments 213–14; gene expression study 331; genome-wide association 330–1; matching and stratifying 233; in mixed methods research 414–17; multi-site, in fMRI 351; quasi-experimental 224–35 (*see also* observational studies); regression discontinuity 234; in safety research 381–2, 393
 differential equation models 172–84
 differential item functioning (DIF) 73–4
 difficulty, item 60–1, 62
 DIKW *see* data to wisdom model
 discretization 300–2
 discretization class attribute interdependency maximization (CAIM) 301–2
 dissemination 9–10, 22–3, 423–30; in economic evaluations 408–9; in mixed methods research 418
 distal outcome 142
 DNA 324, 332
 domain knowledge 294
 dose 246
 dynamical systems 170–3
 dynamical systems analysis 170–84; derivatives 172–3; differential equation models 172–84; exact discrete model (EDM) 183; first-order
- differential equation model 174–80; latent derivative estimates (LDE) 182–3; model estimation 181–3; model selection 177–81; observed derivative estimates 182; stochastic innovations 174–7

 ecological momentary assessment (EMA) 92
 effect modification 240–1
 effect size 209; in meta-analysis 370–1
 efficacy trial 199
 electric and electromagnetic measurement 106–8
 electronic health records 269, 271–2, 280
 embedded mixed methods design 416
 Empirical Bayes estimates 163–4
 endogenous variable 119
 energy principles 98–9
 energy transfer 99
 English, in dissemination 424, 428–9
 enthalpy 99
 entropy (energy transfer) 99
 entropy (latent class) 144
 environment 17
 environmental sensors 92
 epigenomics 331–2
 EQS 131–2
 equivalent models 125
 estimation: consistency 70; general growth mixture models 143; IRT 66–71; joint models 256; SEM 125–6
 ethics 22, 47, 210, 272, 286–7; in safety research 390–1
 event 17, 20, 87, 253; processes 256; in safety research 382–4
 evidence: practice-based 272–3
 evidence-based practice 8–9, 22–3, 196–7, 361–3, 425–7
 evidence networks 375–6
 exact discrete model (EDM) 183
 exercise and physical activity outcome measurement 101–2
 exogenous variable 119
 experimental design 19, 213–14; and mixed methods 413–14; in non-randomized studies 219, 224–235; in pilot studies 199–204; in safety research 381–2
 explanatory mixed methods design 416
 exploratory mixed methods design 416–17
 exposome 17
 external validity 226

 factor analysis 122–3
 feasibility 200
 feature 298; feature selection 298–9
 fidelity 194–5, 246
 first-order differential equation model 174–80
 fit 128–9, 143
 fixed effects 160–6

- flow of body fluids measurement 109–12
- fMRI 336–57
- force and pressure measurement 108–9
- forest plot 372–3
- gene expression study 331
- general growth mixture models 130, 137–51
- generalizability 226
- generalized least squares (GLS) estimator 126
- generalized local linear approximation (GLLA) 182
- generalized orthogonal derivative estimates (GOLD) 182
- genome 324–5
- genome-wide association study 330–1
- Gibbs free energy 99
- goodness of fit: in data mining 313–16; in general
 - growth mixture models 143–4; in IRT 73–4;
 - in SEM 128–9
- greatest lower bound (glb) reliability 33–4
- guessing parameter 63
- Guyatt's responsiveness ratio (GRR) 52
- health 3–4, 15–16, 20, 40, 324; global 11, 427
- health trajectory 21
- hierarchical data 155
- hierarchical models 154
- hopelessness trajectory model 138
- indirect causal effect 243–4
- individual growth parameters 140
- inference 217–18
- inferential threats in quasi-experimental designs 226–7
- informatics agenda 290
- information 19, 269, 271–2; in item response theory 72–3
- instrumental variable 238, 247–9; rarity 248
- intercept-only multilevel model 159–60
- internal validity 226; threats 226–7
- interrater agreement and reliability 94–5
- intervention: definition 189, 190–1; design 192–4;
 - evaluation 195–6; mediators 192; moderators 192; operationalization 194–5; outcomes 191;
 - for safety 385–6, 393; theory 189–92
- interventions 190–1; in mixed methods research 413–14
- intraclass correlation 160
- item characteristic curve (ICC) 60–6
- item response theory (IRT) 58–76
- item response theory models 59–66; 1-PL
 - model for dichotomous items 60–1; 2-PL
 - model for dichotomous items 61–2; 3-PL
 - model for dichotomous items 62–3; graded response model for polytomous items 63–6;
 - item characteristic curve (ICC) 60–6; item information and test information 72–3; model
 - fit 73–4; model parameter estimation 66–71;
 - software for estimation 74
- knowledge 269–70; discovery in databases 285, 295–7; nurses as knowledge workers 276
- known groups 143
- lambda2 reliability estimate 34
- latent classes 137, 149–51; in joint models 261
- latent derivative estimates (LDE) 182–3
- latent growth curve models 130, 137
- latent subgroups 137
- latent trait models *see* item response theory models
- latent variable 119
- latent variable mixture models 130
- learning health system 289–90
- level 1 and level 2 units 155
- limits of agreement 52–3
- LISREL 121, 131–2; matrix notation 121
- local linear approximation (LLA) 182
- longitudinal design 228–32, 336; and responsiveness 52; *see also* time series
- machine learning 294; one-class 304, 306;
 - rule-based 303, 306–13; supervised 303–6;
 - unsupervised 303
- magnetic resonance imaging 107; *see also* fMRI
- manifest variables 119
- markers recurrent 261
- matching 233
- maximum likelihood (ML) estimator 126
- mean structures 130
- measured variables 119
- measurement: of basal metabolic rate 100; of
 - exercise and physical activity 101–2; of force and pressure 108–9; level of 46; precision 37–9, 51–3, 71–2; procedures 45–7, 89–93; of temperature 102–3
- measurement theory 19, 29–41, 47–53, 58–77, 94–5
- mechanical energy 101–2
- mediator variables 192, 238, 242–5
- meta-analysis 369–75
- metabolic equivalents (METs) 101
- meta-paradigm 16–17; biopsychosocial perspective 6–7
- minimally important change (MIC) or difference (MID) 52–3
- misclassification matrix 315–17
- missing data 144, 220
- mixed methods 411–19
- mixed models 154
- mixtures of distributions 130, 137
- model: exact discrete model (EDM) 183; joint 253–63
- model(ing) 20–1, 297, 302

- model assessment 126–8
- model estimation 125–6, 143, 181–3
- model fit 128, 143
- model generating 128
- model identification 125
- model selection 127, 143–4, 177–81
- model testing 127, 73–4
- moderator variables 192, 238, 239–41; in
 - dynamical systems 178–80; in meta-analysis 370, 372–4; in safety research 389–90
- Mplus* 131–2
- multilevel model file structure 158–9
- multilevel models 154–69
- multiple group models 130
- natural indirect effect 243–4
- nested data 129, 155
- non-randomized studies 238–49; *see also*
 - observational studies; quasi-experimental designs
- non-recursive path model 123
- nuisance regressors in fMRI 348, 349
- nursing science 1–2, 3–13, 15; and biology 6–7
- nursing theory 15–23, 275
- observational studies 217–19; *see also* quasi-experimental designs
- observed derivative estimates 182
- observed variable path model 120, 123–4
- OpenMx 131, 183
- parallel tests 31
- parameter 119
- parameter invariance 59
- path analysis model 120, 123–4
- path diagram 119–21
- patient-reported outcomes 45–6, 58
- PCORI 397
- person 17
- Phase III trial 199
- phenotype 326
- PICO 364, 366
- pilot studies 199–210; blinding 205–6; design 200–7; effect size 209; efficacy trials 199; ethical issues 210; feasibility and acceptability 200; outcome measurement 206–7; participants 201–5, 207; randomization 205; randomized clinical trial (RCT) 199–210; registration 210; sample size 209
- policy 10–11, 12, 428
- population 21–2
- population heterogeneity 137
- potential outcomes 214–5, 232, 239; *see also* Rubin Causal Model
- practical issues in instrument selection 47
- pragmatism 412
- pressure 108–9
- principal strata effect 243–4
- program fidelity 246
- PROMIS 55, 58, 75
- propensity score 216, 233–4
- PROSPERO 368
- proteomics 332–3
- quality adjusted life years 402–3
- quanta matrix 301
- quasi-experimental designs 224–35; control
 - group designs with no pretest 229–30; control group with pretest designs 230–1; design variations 227–8; difference in difference design 234–5; inferential threats 226–7; matching and stratifying 233; regression discontinuity design 234; single group designs with no control group 228–9; treatment assignment mechanisms 232–3
- random effects models 154
- random intercepts and slopes model with
 - individual level covariates 161–6
- randomization 205–6; feasibility, in safety research 393
- randomized clinical trial (RCT) 199–210, 213–17
- randomized clinical trial (RCT) complications 220
- randomized experiment 213–17
- random measurement error 30
- receiver operating curves 317–18
- recursive path model 123
- registration: of clinical trials 210; of systematic reviews 368
- regression discontinuity design 234
- reliability: alpha 32–5; and standard error of
 - measurement 50–1; definition 31, 50; interrater 94; parallel-test 37; retest 36–7; split-half 35–6
- residuals: Martingale 259; and time trend in joint models 260
- responsiveness 52–3
- resting metabolic rate 100
- R structural equations modeling software 131
- Rubin Causal Model (RCM) 213, 220, 232–3
- safety 381–94
- sample 21–2
- sample size: in GWAS 330; in pilot studies 209–10; in structural equation modeling 130
- sampling variance: difference in means 218; *see also* standard errors
- sensitivity analysis in economic evaluations 407–8
- smallest detectable change (SDC) 52–3
- software: data mining 321; dynamical systems 181–3; economic analyses 407; fMRI 343; genetic association studies 325; IRT 74; joint models 256; Language Environment Analysis (LENA) 91; meta-analysis 375; multilevel models 168–9; nursing terminologies 284; qualitative analysis 417; reliability estimation 34; SEM 131–2; WEKA 302, 306

- sound measurement 103–6
- split-half reliability 35–6
- stable unit-treatment value assumption (SUTVA) 215, 232
- standard error of measurement 37–8, 50–1, 72
- standard errors 21; and conditional independence 69; of the difference 39; in fMRI studies 342, 345, 349; and instrumental variables 248–9; in meta-analysis 370; in nested data 129, 164; in structural equation models 126, 129; *see also* sampling variance
- Standards for Educational and Psychological Testing* 39
- statistical conclusion validity 226
- statistical parametric map in fMRI 341
- stochastic innovations 174–7
- stratifying 233
- structural equation modeling software 131–2
- structural equation models 119–33
- structural equation models with latent variables 124
- survival 253–4
- SUTVA *see* stable unit-treatment value assumption
- systematic review 363; of instruments 53–5
- task 337
- tau (τ) equivalence 33
- temperature 102–3
- temporal correlation in fMRI 340, 352
- temporal rate in fMRI 337
- terminologies 281, 284
- test construction 40–1
- theory: translation to statistical model 119–22, 138, 158–64, 170–2, 177–81; *see also* measurement theory; nursing theory
- theory-based interventions 189–97; in safety research 392–3
- thermal energy and temperature measurement 102–3
- threats in quasi-experiments 226–7
- time 17, 85–6, 240, 242, 253; and behavioral topography 85; in moderator variable models 240
- time sampling 87–8
- time series: in fMRI 339, 345; in joint models 253–4
- time-varying confounding 242
- tokenism in mixed methods research 419
- trajectory 17, 253; and time trend in joint models 260; *see also* health trajectory
- translational science 7–8, 22, 196–7
- treatment adherence 246
- triangulation 415
- true score 30–1
- ultrasound measurement 104–6
- validation, models in data mining 313–15
- validity, test score 39–41, 49–50
- variables 21, 46–8, 59, 129; continuous 119; dichotomous 60–3; latent 119, 137; polytomous 63–6; *see also* discretization, event, mediator variables, moderator variables
- variance components in joint models 258–9
- variance-covariance matrix 120
- visualization 318–21
- voxel 337
- whole interval sampling 89
- wisdom 270, 276