ICSA Book Series in Statistics
Series Editors: Jiahua Chen · Ding-Geng (Din) Chen

Karl E. Peace
Ding-Geng Chen
Sandeep Menon *Editors*

Biopharmaceutical Applied Statistics Symposium

Volume 2 Biostatistical Analysis of Clinical Trials

Springer

ICSA Book Series in Statistics

More information about this series at http://www.springer.com/series/13402

Karl E. Peace · Ding-Geng Chen
Sandeep Menon
Editors

Biopharmaceutical Applied Statistics Symposium

Volume 2 Biostatistical Analysis of Clinical Trials

Springer

Editors
Karl E. Peace
Jiann-Ping Hsu College of Public Health
Georgia Southern University
Statesboro, GA, USA

Sandeep Menon
Boston University
Cambridge, MA, USA

Ding-Geng Chen
School of Social Work & Gillings School of Global Public Health
University of North Carolina
Chapel Hill, NC, USA

and

University of Pretoria
Pretoria, South Africa

ISSN 2199-0980 ISSN 2199-0999 (electronic)
ICSA Book Series in Statistics
ISBN 978-981-10-7825-5 ISBN 978-981-10-7826-2 (eBook)
https://doi.org/10.1007/978-981-10-7826-2

Library of Congress Control Number: 2017964432

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Currently, there are three Volumes of the BASS Book Series, spanning 45 chapters. Chapters in this book are contributed by invited speakers at the annual meetings of the Biopharmaceutical Applied Statistics Symposium (BASS). Volume 1 is titled: Design of Clinical Trials and consists of 15 chapters; Volume 2 is titled Biostatistical Analysis of Clinical Trials and consists of 12 chapters; and Volume 3 is titled Pharmaceutical Applications and consists of 18 chapters. The three volumes include the works of seventy authors or co-authors.

History of BASS: BASS was founded in 1994, by Dr. Karl E. Peace. Dr. Peace is the Georgia Cancer Coalition Distinguished Scholar/Scientist, Professor of Biostatistics, Founding Director of the Center for Biostatistics, and Senior Research Scientist in the Jiann-Ping College of Public Health at Georgia Southern University.

Originally, there were three objectives of BASS. Since the first editor founded the Journal of Biopharmaceutical Statistics (JBS) 3 years before founding BASS, one of the original objectives was to invite BASS Speakers to create papers from their BASS presentations and submit to JBS for review and publication. Ergo, BASS was to be a source of papers submitted to JBS to assist in the growth of the new journal JBS. The additional two objectives were:

- to provide a forum for pharmaceutical and medical researchers and regulators to share timely and pertinent information concerning the application of biostatistics in pharmaceutical environments; and most importantly,
- to provide revenues to support graduate fellowships in biostatistics at the Medical College of Virginia (MCV) and at the Jiann-Ping Hsu College of Public Health at Georgia Southern University (GSU).

After the JBS was on firm footing, the first objective was formally dropped. In addition, the third objective was expanded to include potentially any graduate program in biostatistics in the USA.

BASS I (1994) was held at the Hyatt Regency in Orlando, FL; BASS II–III were held at the Hilton Beach Resort, Inner Harbor, in San Diego, CA; BASS IV–VII were held at the Hilton Oceanfront Resort Hotel, Palmetto Dunes, in Hilton Head Island, SC; BASS VIII–XII were held at the Desoto Hilton; and BASS XIII–XVI were held at the Mulberry Inn, both located in the Historic District of Savannah, GA. BASS XVII was held at the Hilton Resort Hotel at Palmetto Dunes, Hilton Head Island, SC. BASS XVIII–XIX were held at the Mulberry Inn in Savannah. To mark the twentieth Anniversary BASS meeting, BASS XX was held in Orlando at the Hilton Downtown Orlando Hotel. BASS XXI was held at the Holiday Inn Crowne Plaza in Rockville, MD, whereas BASS XXII and XXIII were held at the Radisson Hotel in Rockville, Maryland.

BASS XXIV (www.bassconference.org) was held at the Hotel Indigo in the charming historic Georgia city of Savannah. More than 360 tutorials and 57 1-day or 2-day short courses have been presented at BASS, by the world's leading authorities on applications of biostatistical methods attendant to the research, clinical development, and regulation of biopharmaceutical products. Presenters represent the biopharmaceutical industry, academia, and government, particularly the NIH and FDA.

BASS is regarded as one of the premier conferences in the world. It has served the statistical, biopharmaceutical, and medical research communities for the past 24 years by providing a forum for distinguished researchers and scholars in academia, government agencies, and industries to conduct knowledge sharing, idea exchange, and creative discussions of the most up-to-date innovative research and applications to medical and health care to enhance the health of general public, in addition to providing support for graduate students in their biostatistics studies. Toward this latter end, BASS has provided financial support for 75 students in completing their Master or Doctorate degree in Biostatistics. In addition, BASS has provided numerous travel grants to Doctorate-seeking students in Biostatistics to attend the annual BASS meeting. This provides a unique opportunity for students to broaden their education, particularly in the application of biostatistical design and analysis methods, as well as networking opportunities with biostatisticians from Academia, the Pharmaceutical Industry, and governmental agencies such as the FDA.

Volume II of the BASS Book Series, entitled Biostatistical Analysis of Clinical Trials, consists of 12 chapters. Chapter 1 presents collaborative targeted maximum likelihood estimation methods to assess causal effects in observational studies. Chapter 2 discusses the use of generalized tests in clinical trials. Chapter 3 presents discrete time-to-event and score-based methods with application to a composite endpoint for assessing evidence of disease activity-free. Chapter 4 discusses methods for imputing missing data using a surrogate biomarker and presents an analysis of the incidence of endometrial hyperplasia. Chapter 5 deals with advancing the interpretation of patient-reported outcomes. Chapter 6 provides a primer on network meta-analysis with an example.

Chapter 7 presents methods for detecting safety signals among adverse events reported in clinical trials. Chapter 8 provides methods for meta-analysis for rare events reported in clinical trials. Chapter 9 provides a treatise on missing data.

Chapter 10 discusses Bayesian subgroup analysis using hierarchical models. Chapter 11 presents a question-based approach to the analysis of safety data. Finally, in Chap. 12, the analysis of two-stage adaptive seamless clinical trial design is presented.

We are indebted to all the presenters, program committee, attendees, and volunteers who have contributed to the phenomenal success of BASS over its first 24 years, and to the publisher for expressing interest in and publishing the Series.

Statesboro, USA
Karl E. Peace, Ph.D.
Jiann-Ping Hsu College of Public Health
Georgia Southern Univesity

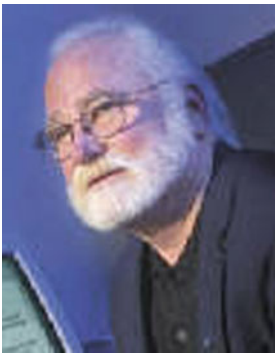
Chapel Hill, USA/Pretoria, South Africa
Ding-Geng Chen, Ph.D.
Professor, University of North Carolina
Extraordinary Professor, University of Pretoria

Cambridge, USA
Sandeep Menon
Vice President and Head of Early
Clinical Development, Biostatistics

Contents

Editors and Contributors

About the Editors



Prof. Karl E. Peace is currently Professor of Biostatistics, Senior Research Scientist, and Georgia Cancer Coalition Distinguished Scholar, in the Jiann-Ping Hsu College of Public Health (JPHCOPH) at Georgia Southern University, Statesboro, GA. He is a Fellow of the American Statistical Association (ASA), the Founding Editor of the Journal of Biopharmaceutical Statistics, Founding Director of the Center for Biostatistics in the JPHCOPH, Founder of BASS, and the Endower of JPHCOPH. He is the recipient of numerous awards and citations from the ASA, the Drug Information Association, the Philippine Statistical Association, BASS, and government bodies. He was cited by US and State of Georgia Houses of Representatives and the House of Delegates of Virginia for his contributions to Education, Public Health, Biostatistics, and Drug Research and Development. He is the author or editor of 12 books and over 100 publications.

Prof. Ding-Geng Chen is a Fellow of the American Statistical Association and currently the Wallace H. Kuralt Distinguished Professor at the University of North Carolina at Chapel Hill, USA, and an extraordinary Professor at University of Pretoria, South Africa. He was a Professor at the University of Rochester and the Karl E. Peace endowed eminent scholar chair in Biostatistics at the Department of Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University. He is also a Senior Consultant for biopharmaceuticals and government agencies with extensive expertise in clinical trial biostatistics and public health statistics. He has written more than 150 refereed publications and co-authored/co-edited 12 books on clinical trial methodology, meta-analysis, causal inference, and public health statistics.

Dr. Sandeep Menon is currently the Vice President and the Head of Early Clinical Development Statistics at Pfizer Inc. and also holds adjunct faculty positions at Boston University, Tufts University School of Medicine and Indian Institute of Management (IIM). He is the elected fellow of American Statistical Association. He is internationally known for his technical expertise especially in the area of adaptive designs, personalized medicine, multiregional trials, and small populations. He has co-authored and co-edited books and contributed to influential papers in this area. He is the Vice Chair of Cross Industry/FDA-Adaptive Design Scientific Working Group under DIA (Drug Information Association); in the program committee for BASS and ISBS; and is in the advisory board for the M.S. in Biostatistics program at Boston University. He is serving as an Associate Editor of American Statistical Association (ASA) journal *Statistics in Biopharmaceutical Research* (SBR) and as a selection committee member of *Samuel S. Wilks Memorial Award* offered by ASA.

Contributors

William L. Baker University of Connecticut, Storrs, CT, USA

Andrew G. Bushmakin Pfizer Inc., Groton, CT, USA

Joseph C. Cappelleri Pfizer Inc., Groton, CT, USA

Ding-Geng Chen Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA; Department of Statistics, University of Pretoria, Pretoria, South Africa

Shein-Chung Chow Duke University School of Medicine, Durham, NC, USA

Jared C. Christensen Early Clinical Development, Pfizer Inc., Cambridge, MA, USA

Steven A. Gilbert Early Clinical Development, Pfizer Inc., Cambridge, MA, USA

Susan Gruber Cambridge, MA, USA

H. Jiang Janssen Research & Development, LLC, Raritan, NJ, USA

P. Lim Janssen Research & Development, LLC, Titusville, NJ, USA

Min Lin Food and Drug Administration, Silver Spring, MD, USA

Melvin S. Munsaka Gurnee, IL, USA

Stephan Ogenstad Statogen Consulting LLC, Wake Forest, NC, USA

Macaulay Okwuokenye Jiann-Ping Hsu College of Public Health, Georgia Southern University and University of New England, Syros Pharmaceutical, Cambridge, MA, USA

Karl E. Peace Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, USA

Gene Pennello Division of Biostatistics, Food and Drug Administration, Center for Devices and Radiological Health, Silver Spring, MD, USA

Mark Rothmann Division of Biometrics II, Food and Drug Administration, Center for Drug Evaluation and Research, Silver Spring, MD, USA

Mark van der Laan Jiann-PingHsu/Karl E. Peace Professor in Biostatistics & Statistics, University of California at Berkeley, Berkeley, CA, USA

Richard C. Zink TARGET PharmaSolutions, Chapel Hill, NC, USA; Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Chapter 1
Collaborative Targeted Maximum Likelihood Estimation to Assess Causal Effects in Observational Studies

Susan Gruber and Mark van der Laan

1.1 Introduction

Randomized controlled trials (RCT) and observational studies (OS) are carried out to address a wide variety of problems in medicine and public health. RCTs are considered a strong source of evidence of causal associations because treatment is randomized (Burns et al. 2011). However, if there is informative loss to follow-up or lack of adherence, an unadjusted analysis of RCT data will be biased. Furthermore, for some questions an RCT is not an ethical or feasible option and only observational data are available. OS are common tools for post-market drug safety monitoring and assessing the impact of environmental exposures on chronic disease. A naive analysis that ignores selection bias and other sources of confounding will produce a biased estimate of the causal effect. Targeted learning (TL) provides a framework for using data to answer these kinds of questions (van der Laan and Rose 2011).

In the TL paradigm study data are viewed as realizations from an underlying joint distribution of the data, P_0. Common effect measures of interest such as the hazard ratio (HR), odds ratio (OR), relative risk (RR), and additive treatment effects (ATE) correspond to parameters of P_0. Although P_0 is unknown, when causal assumptions are met these parameters can be estimated from data. The goal of TL is to construct an estimator that is maximally precise and minimally biased. TL uses two core methodologies, super learning (SL) and targeted minimum loss-based estimation (TMLE) (van der Laan and Rubin 2006; van der Laan et al. 2007; van der Laan and Rose 2011). SL is an ensemble machine learning algorithm for prediction that provides flexible, data-adaptive modeling. SL avoids imposing unwarranted parametric

S. Gruber (✉)
Putnam Data Sciences, LLC, 85 Putnam Avenue, Cambridge, MA 02139, USA
e-mail: sgruber@putnamds.com

M. van der Laan
School of Public Health, University of California at Berkeley, 108 Haviland Hall, Berkeley, CA 94720-7360, USA
e-mail: laan@berkeley.edu

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_1

assumptions upon the distribution of the data that can introduce bias. TMLE is a semi-parametric efficient double-robust (DR) substitution estimator that tailors the estimation procedure to the target parameter of interest. Unlike maximum likelihood estimators that optimize a global likelihood, TMLE aims to make a more favorable bias/variance trade off with respect to the target statistical estimand. Collaborative TMLE (C-TMLE) is an extension of TMLE designed to further improve this trade off when there is a sparsity of information in the data for estimating the target parameter (van der Laan and Gruber 2010; Gruber and van der Laan 2010b, 2015).

As a running example consider the problem of estimating the ATE of a binary exposure, A, on a binary outcome, Y, adjusting for covariates W, using observational data. Many propensity score (PS)-based methods and outcome regression modeling approaches to adjusting for confounding exist. PS-based estimators including matching, stratification, and inverse probability weighting (IPW), make use of associations between covariates and observed treatment (Rosenbaum and Rubin 1983; Davidian and Lunceford 2004; Hernan et al. 2000). Outcome regression modeling adjusts for bias in the estimated association between treatment and the outcome by including potential confounders in the model. A DR estimator such as TMLE combines these two strategies, and is consistent if either the PS or the outcome regression model is correctly specified.

The literature on PS-based estimators recognizes that the choice of covariates to include in the PS model affects efficiency, and that model misspecification can amplify bias (Caliendo and Kopeinig 2008; Sekhon 2008; Petersen et al. 2010; Pearl 2011). These problems are exacerbated when within some strata of W very few treated (or untreated) subjects are available to inform the effect estimate. In other words, when combinations of covariates are highly correlated with treatment it becomes difficult to tease apart the causal effect of treatment versus the effect of the covariates themselves. This lack of robust common support is known as poor overlap in the matching literature. In the causal inference literature it is termed a near violation of the positivity assumption that $0 < P(A = 1 \mid W) < 1$. Outcome regression-based effect estimates are also impacted. They are primarily driven by untestable modeling assumptions rather than by data.

One response to sparsity involves re-framing the question by defining an alternative target parameter. For example, one might opt to estimate the effect of treatment among the treated (ATT), rather than among the entire study population. Another way of re-defining the target population is to trim observations where the probability of receiving or not receiving treatment is close to 0 or 1, under the assumption that the treatment effect is of interest only in the population where realistic treatment options exist (Smith and Todd 2005). However, interpretability of this parameter suffers from a limited ability to characterize the study population (King et al. 2014).

Although DR estimators are not immune to near violations of the positivity assumption, TMLE dampens finite sample bias and variability by virtue of being a bounded substitution estimator (Kang and Schafer 2007b; Robins et al. 2007; Porter et al. 2011). In addition, theoretical results indicate that when the outcome regression model is partially informative there is opportunity for a DR estimator to further reduce mean squared error (MSE) in these sparse data settings (van der

Laan and Gruber 2010; Gruber and van der Laan 2010a). This insight motivated the development of the C-TMLE.

C-TMLE is an extension of TMLE that data adaptively estimate the propensity score model in response to bias inadequately addressed by the outcome model. The remainder of this chapter describes the underlying principles of C-TMLE and provides a general template for a C-TMLE. Several implementations of C-TMLE that have been discussed in the literature are also presented, along with applications simulated and real-world datasets. We begin with a brief review of TMLE.

1.2 Targeted Minimum Loss-Based Estimation

TMLE is a methodology for estimating any pathwise differentiable parameter of a probability distribution, including common causal effect parameters (van der Laan and Rubin 2006; van der Laan and Rose 2011). The parameter of interest is a feature of the joint distribution of the data. It is defined as a mapping from the class of probability distributions under consideration, $\mathscr{M}$, to the parameter space, $\Psi: \mathscr{M} \rightarrow \mathbb{R}$. The target parameter is evaluated by applying the mapping to the true data distribution, $\psi_0 = \Psi(P_0)$. (In the conventions of the TMLE literature, a '0' subscript indicates a true value and the subscript 'n' denotes an estimate.) As a substitution estimator, the TMLE is evaluated by applying the mapping to an estimate of P_0, $\psi_n = \Psi(P_n)$. When the data distribution factorizes into a Q component required for evaluating the mapping, and a nuisance component g, we can also write $\psi_0 = \Psi(Q_0)$.

The TMLE procedure is carried out in two stages. In Stage 1 an initial estimate of Q_0 is obtained. If this estimated Q_n^0 is not consistent for Q_0, then $\Psi(Q_n^0)$ produces a biased estimate of ψ_0. Stage 2 provides an opportunity to use information in g to reduce any remaining bias for ψ_n in Q_n^0. This is accomplished by defining a parametric submodel with fluctuation parameter, ε. The choice of submodel and procedure for fitting ε are judiciously crafted to ensure that the closure of the linear span of score equations solved when fitting ε by maximum likelihood includes the efficient influence curve estimating equation, and that the updated estimate, Q_n^* remains within the statistical model. We illustrate these general concepts within the context of estimating the ATE.

Using TMLE to Estimate the ATE

Consider the unobservable full data $X^{full} = (W, Y_0, Y_1) \sim P_0$, where W is a vector of baseline covariates and Y_0 and Y_1 correspond to potential (or counterfactual) outcomes that would occur under no treatment and treatment, respectively. The ATE parameter is defined in terms of this full data as $\mathrm{E}_0(Y_1 - Y_0)$. However, X^{full} is unobservable because only one of the potential outcomes actually occurs in the real world.

Coarsening is the process that gives rise to the missing data structure of the observed data, in which only the potential outcome corresponding to treatment A at level $a = 0$ or 1 is recorded. The data are coarsened at random (CAR) when the treatment assignment mechanism that gives rise to the missingness is a function of

observed covariates only. Observed data are a coarsened version of the full data in which binary treatment indicator A indicates which potential outcome is observed. The observed data consists of n independently and identically distributed observations $O = (W, A, Y) \sim P_0$. When causal assumptions are met $Y = Y_a$ and the corresponding statistical parameter is given by the mapping, $\psi_0 = \Psi(P_0) = \mathrm{E}_0[\mathrm{E}_0(Y \mid A = 1, W) - \mathrm{E}_0(Y \mid A = 0, W)]$.

An efficient DR estimator solves an estimating equation based in the efficient influence curve of the pathwise derivative of the target parameter, $D^*(P)$. A DR estimator is consistent in CAR data structures when either Q_0 or g_0 is consistently estimated, and maximally efficient when both are correct (van der Laan and Robins 2003). For parameters where $D^*(P)$ allows an estimating function representation $D^*(\psi, \eta)$, it is possible to directly solve $P_n D^*(\psi, \eta) = 0$. The augmented-IPW estimator takes this approach (Robins and Rotnitzky 1995, 2001; Robins 2000). In contrast, although TMLE also ensures the efficient influence curve estimating equation is solved, i.e., that it satisfies $P_n D^*(Q_n^*, g_n) = 0$, the TMLE is defined as a substitution or plug-in estimator, $\psi_n = \Psi(P_n)$. TMLE can therefore be applied to estimate parameters where no estimating function representation of $D^*(P)$ exists.

For the ATE parameter the Q component of the distribution factorizes as $Q_0 = (Q_{0Y}, Q_{0W})$, the distributions of Y and of W, respectively. The empirical distribution of W provides a consistent non-parametric estimate of Q_{0W}. With regard to Q_{0Y} we note that the ATE mapping only requires estimates of the the conditional mean outcome, $\bar{Q}_0(A, W) \equiv \mathrm{E}_0(Y \mid A, W)$, not the entire density. Stage 1 of TMLE focuses on correctly modeling the outcome regression. Data-adaptive SL is the recommended approach to producing the initial estimate, $\bar{Q}_n^0(A, W)$. We recognize that even when SL is used if $\bar{Q}_n^0$ does not converge to Q_0 or the rate of convergence is slow, then there may be residual bias in $\psi_n = \Psi(\bar{Q}_n^0)$. Stage 2 of TMLE provides an opportunity to reduce this bias.

In Stage 2 TMLE uses information in g to fluctuate the estimate of $\bar{Q}_0$. In this data structure g refers to the conditional distribution of A given W. The goal of this targeting step is to modify $\bar{Q}_n^0$ in a way that improves estimation of ψ. The key is crafting a parametric submodel with a fluctuation covariate $H_g(A, W)$ designed so that when the ε-parameter of the submodel is fit by maximum likelihood, the closure of the linear span of the generated score equations includes the efficient influence curve estimating equation. This approach guarantees that (Q_n^*, g_n) solves $P_n D^*(Q_n^*, g_n) = 0$, where $\bar{Q}_n^*$ is the targeted update of $\bar{Q}_n^0$.

The efficient influence function for the ATE parameter is given by

$$D^*(P)(O) = H(A, W)[Y - \bar{Q}(A, W)] + \bar{Q}(1, W) - \bar{Q}(0, W) - \psi,$$

with $H(A, W) = \frac{A}{g(1,W)} - \frac{1-A}{1-g(1,W)}$ (van der Laan and Rubin 2006). Our goal of ensuring that maximizing the likelihood when fitting ε ensures solving the empirical efficient influence function equation is satisfied by defining the submodel

$$\mathrm{logit}\left[\bar{Q}_n^*(A, W)\right] = \mathrm{logit}\left[\bar{Q}_n^0(A, W)\right] + \varepsilon H(A, W).$$

The updated $\bar{Q}_n^*$ is given by setting

$$\text{logit}[\bar{Q}_n^*(0, W)] = \text{logit}[\bar{Q}_n^0(0, W)] - \varepsilon_n/[1 - g_n(1, W)], \tag{1.1}$$

$$\text{logit}[\bar{Q}_n^*(1, W)] = \text{logit}[\bar{Q}_n^0(1, W)] + \varepsilon_n/g_n(1, W). \tag{1.2}$$

Both potential outcomes are evaluated for all subjects. The parameter estimate is calculated as $\psi_n = \frac{1}{n}\sum_{i=1}^{n}\left[\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)\right]$. Note that both $g(1, W)$ and $g(0, W) = 1 - g(1, W)$ are in the denominator of terms in Eqs. 1.1 and 1.2. The positivity assumption ensures that $g(1, W)$ is bounded away from both one and zero in all strata of W. In an ideal RCT where randomization probabilities are known, using g_0 in Stage 2 guarantees consistency of the TMLE. In an OS where g_0 is not known SL can be used to obtain an estimate from data.

1.2.1 Inference

Influence-curve based inference is available for all TMLEs. The asymptotic variance of the estimator is given by $\sigma_\psi^2 = \sigma^2(D^*(P_0))/n$, where $\sigma^2(D^*(P_n^*))$ is the variance of the influence function and n is the number of independent units of observation. This can be used to construct a test statistic for Wald-type hypothesis testing, $T = (\psi_n - \mu)/\hat{\sigma}_\psi$, where μ is the hypothesized parameter value. Double-sided $1 - \alpha$-level confidence intervals are given by $\psi_n \pm Z_{\alpha/2}\hat{\sigma}$. When P_n^* is consistent, $\sigma^2(D^*(P_0))/n$ is an unbiased estimator of the true variance. Otherwise, as long as g_n is consistent inference is conservative.

1.3 C-TMLE

A near violation of the positivity assumption signals a sparsity of information in the data for identifying the target parameter. This can be detected from data by examining whether $g(1, W)$ is close to zero or one within one or more strata of W. Estimates from sparse data can be biased and highly variable, even under correct model specification (Freedman and Berk 2008). C-TMLE addresses this problem by fine tuning the nuisance parameter estimation procedure. The ultimate goal is to minimize mean squared error in the target parameter estimate.

We saw earlier that modeling g within the standard TMLE procedure is independent of how well Q_n^0 approximates Q_0. In contrast, C-TMLE data-adaptively selects a set of confounders to adjust for when modeling g in response to residual bias in $\Psi(Q_n^0)$. Depending on the consistency and rate of convergence of Q_n^0 it may be possible to condition on fewer covariates in the propensity score model than would be required to estimate the full g_0. Conditioning on less will tend to yield propensity scores that are further bounded away from zero and one. Incorporating these less

extreme propensity scores into $H(A, W)$ can reduce finite sample bias and variance in the target parameter estimate.

1.3.1 Collaborative Double Robustness

Theory teaches that when g_0 is used to target any initial Q_n^0 the TMLE will be consistent. Beyond that, there may exist one or more g estimators that condition on less than the full g_0, yet provide consistent estimation of ψ_0 when used to target a misspecified Q_n^0. This *collaborative* double robustness property states that given a limit Q of Q_n, there exists a set of possible limits g of g_n for which the estimator satisfying $P_n D^*(Q_n, g_n, \psi_n) = 0$ remains consistent for ψ_0 (van der Laan and Gruber 2010). Let $\mathcal{G}(Q_n, P_0)$ be the set of g estimators satisfying this condition. Traditional double robustness result tells us that when Q_n^0 is consistent for Q_0 $\mathcal{G}(Q_n^0, P_0)$ contains all conditional distributions of g. At the other extreme, when Q_n^0 is completely uninformative the residual bias in $\psi_n = \Psi(Q_n^0)$ is equal to all of the bias in the crude parameter estimate. In this case any g in $\mathcal{G}(Q_n^0, P_0)$ must condition on a set of covariates sufficient to control for all confounding, and may condition on more. The important point is that for a Q with $\Psi(Q) = \psi_0$ the set $\mathcal{G}(Q, P_0)$ is defined as all g in $\mathcal{G}$ such that $P_0 D^*(Q, g) = 0$. When an estimating function representation exists, this is equivalent to all g for which $P_0 D^*(Q, g, \psi_0) = 0$. If Q_n converges to Q and g_n converges to a g in $\mathcal{G}(Q, P_0)$ then the solution ψ_n of $P_0 D^*(Q_n, g_n, \psi_n)$ will be consistent.

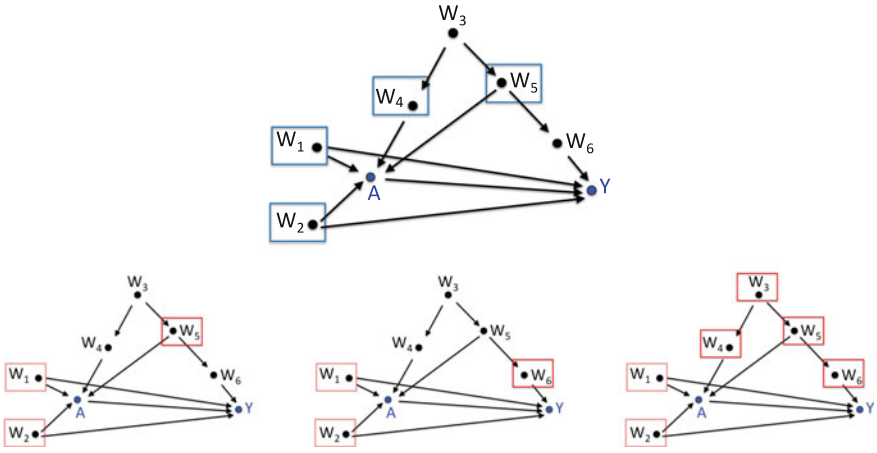


Fig. 1.1 Directed acyclic graph (DAG) with arrows depicting the true causal relationships among covariates W, treatment A, and outcome Y. The true treatment mechanism, g_0 conditions on nodes within boxes (top). Three copies of the same DAG have boxes around covariates that constitute alternate sufficient adjustment sets (bottom)

In addition to the true g_0 there might exist several sufficient g estimators that condition on different adjustment sets. For example, the directed acyclic graph (DAG) in Fig. 1.1 shows the true causal relationships among covariates $W_1, \ldots, W_5$, treatment A, and outcome Y. Common ancestors of A and Y confound the effect of A on Y. The covariates $\tilde{W} = (W_1, W_2, W_4, W_5)$ in boxes at the top of Fig. 1.1 are direct parents of treatment node A. The true $g_0 = P(A = 1 \mid \tilde{W})$. Even when Q_n^0 is uninformative, a DR estimator based on this g_0 is consistent for ψ. Further examination of the DAG suggest other subsets of W that are sufficient to control for all confounding. These include $\tilde{W}' = \{W_1, W_2, W_5\}$, $\tilde{W}'' = \{W_1, W_2, W_6\}, \ldots, \tilde{W}''' = \{W_1, W_2, W_3, W_4, W_5, W_6\}$. Despite the fact that that $g(A, \tilde{W}'')=P(A=1 \mid W_1, W_2, W_6)$ is not a consistent estimator of g_0, $\left\{g(A, \tilde{W}), g(A, \tilde{W}'), g(A, \tilde{W}'), g(A, \tilde{W}')\right\}$ are all sufficient g estimators, and are members of $\mathcal{G}(Q_n^0, P_0)$.

There is a middle ground between the two extremes of a correctly specified initial estimate and a completely uninformative Q_n^0. Consider a $Q_n^{0\prime}$ that adjusts for some, but not all, asymptotic bias. The set of g estimators that in collaboration with $\mathcal{G}(Q_n^{0\prime}, P_0)$ provide consistent estimation of ψ may be larger than the set of sufficient g estimators for the uninformative $Q_n^0, \mathcal{G}(Q_n^0, P_0)$. When $Q_n^{0\prime}$ is partially informative, the residual bias is somewhat less than the full bias in the crude estimate. As long as g conditions on at least $(Q_n^0 - Q_0(0, W), Q_n^0 - Q_0(1, W))$, then in our example conditioning on only a subset of confounders in the propensity score model is sufficient for full asymptotic bias reduction. Collaborative double robustness is a property of all DR estimators including TMLE.

1.3.2 Guiding Principles

Collaborative double robustness teaches us that there are certain circumstances where a DR estimator relying on a misspecified outcome regression model can be consistent for ψ even when the g estimator conditions on less than the full g_0. For a given Q_n^0, the best finite sample adjustment set for estimating g depends upon the underlying causal structure and characteristics of the data. This implies there is an opportunity to reduce MSE in sparse data settings, and motivated the development of C-TMLE.

The main guiding principle for a C-TMLE is that it data-adaptively generates a $g \in \mathcal{G}$ that conditions on a necessary subset of confounders $\tilde{W} \subseteq W$ that explains the difference $(Q_0 - Q_n^0)$. The C-TMLE procedure consists of iteratively constructing a sequence of TMLES $(Q_{n,k}^*, g_{n,k})$ for which the fit for both $Q_{n,k}$ and $g_{n,k}$ is increasing in k. The sequence or g estimators ends with the most nonparametric estimator, the one would use in a standard TMLE. In addition, $Q_{n,k}^*$ is the TMLE using $g_{n,k}$ in the targeting step to update either Q_n^0 itself, or possibly one of the previous TMLEs in the sequence, $Q_{n,j}^*, j < k$. Finally, the criterion for selecting $g_{n,k+1}$, given the past k in the sequence is the increase in fit for Q_0 that occurs during the TMLE update step when using $g_{n,k+1}$ versus $g_{n,k}$.

For intuition consider constructing a sequence of models for g that increase in likelihood and converge towards the true g_0. For example, if the true $g_0 \equiv P(A = 1 \mid W_1, W_2, W_3)$, and W_3 is the strongest predictor of A then sequences s_1 and s_2 both satisfy this requirement.

$$s_1 = \{(W_1), (W_1, W_2), (W_1, W_2, W_3)\},$$
$$s_2 = \{(W_1), (W_3), (W_2, W_3), (W_1, W_2, W_3)\}.$$

However, we also impose an additional constraint that as the model for g grows towards g_0, the likelihood for Q must also be increasing. This provides a meaningful criterion, because it guarantees that $(Q_0 - Q^*_{n,k})$ is shrinking as k increases. By extension, bias in $\Psi(Q^*_{n,k})$ decreases as k increases.

Given a sequence of g estimators we can calculate the corresponding fluctuation covariate $H_1(A, W), \ldots, H_K(A, W)$ used in Stage 2 of the TMLE procedure, then evaluate K corresponding updates of the initial Q^0_n. This produces a sequence of targeted estimates, $(Q^*_{n,1}, \ldots, Q^*_{n,K})$. Depending on the amount of residual bias, an early model for g might be inadequate, while later models might needlessly increase variance without providing a commensurate reduction in bias. Since our main concern is unbiased estimation of ψ, the C-TMLE is the candidate TMLE in the sequence that minimizes cross-validated loss with respect to Q. We define the estimator mapping $\hat{Q}^*_k(P_n)$ as the k-th TMLE in the sequence applied to data P_n. The cross-validated risk of this candidate estimator with respect to loss function $\mathscr{L}(Q)$ is as follows,

$$\text{cv-Risk}_n(k) = \frac{1}{V} \sum_{v=1}^{V} P^1_{n,v} \mathscr{L}(\hat{Q}_k(P_{n,v})),$$

where $P^1_{n,v}$ is the validation sample and $P_{n,v}$ is the training sample for the v-th split according to the V-fold cross validation scheme. A concave function such as the negative log likelihood is a convenient choice of loss function, but any function that is minimized at the true Q_0 is a valid loss function.

C-TMLE Template

Stage 1. Obtain an initial estimate Q^0_n of Q_0.
Stage 2. Target the initial estimate using a data-adaptively estimated g
1. Construct a sequence of nested g estimators and corresponding TMLEs based on updates to Q^0_n, $\{(g_{n,1}, Q^*_{n,1}), \ldots, (g_{n,K}, Q^*_{n,K})\}$.
2. Identify Q^*_{n,k^*}, the candidate TMLE in the sequence that minimizes the cross validated risk, cv-Risk$_n(k)$.

Evaluate the parameter estimate: $\psi_n = \Psi(Q^*_{n,k^*})$.

Note that although the additional constraint on the sequence of models for g reduces the number of permissible sequences, it does not uniquely identify an optimal sequence. More and less aggressive approaches to constructing the sequence

of g estimators give rise to a family of C-TMLEs, each of which adheres to a general template. The next section describes several implementations of C-TMLE that follow an overall strategy of creating a sequence of nested models for g, corresponding candidate TMLEs, and using cross-validation to select the best candidate in the sequence. Each C-TMLE algorithm is motivated by the fact that TMLE is a substitution estimator, $\psi_n = \Psi(Q_n^*)$, thus all aspects of the procedure can be guided by the goodness of fit for Q.

1.3.3 Inference

Because the sequence of g-estimators is finite, with probability tending to 1 the C-TMLE will asymptotically select the final $(g_{n,K}, Q_{n,K}^*)$ at the cross-validation step. As a consequence, the C-TMLE represents a finite sample adjustment of the TMLE and is asymptotically equivalent with this TMLE. Thus, inference as described earlier in Sect. 1.2.1 for TMLE remains valid for C-TMLE.

C-TMLE can at times be super efficient. Super efficiency will occur when the initial Q_n^0 is the MLE for a correct parametric model, in the case where g_n converges to a stable limit. In general super efficiency will only occur on a set of data distributions with measure zero. If the length of the sequence of g estimators were allowed to increase with sample size rather than be set to a fixed K then an analysis of C-TMLE would be more involved.

1.4 C-TMLE Algorithms

1.4.1 Greedy C-TMLE

The first C-TMLE introduced in the literature used a greedy targeted forward selection algorithm to construct a nested sequence of models for g in order to estimate the ATE parameter (Gruber and van der Laan 2010a). As summarized in Algorithm 1, at each step a single additional covariate is incorporated into the model for g, until the final model that conditions on all K available covariates. This indexes a corresponding sequence of TMLEs, $\{\bar{Q}_{n,1}^*, \ldots \bar{Q}_{n,k}^*\}$. The C-TMLE is evaluated as $\psi_n = \Psi(\bar{Q}_{n,k^*}^*)$, where k^* corresponds to the k for which $\bar{Q}_{n,k}^*$ minimizes $\mathscr{L}(\bar{Q})$. The initial article presented a penalized loss function, where the likelihood for Q was penalized by an estimated bias and variance term that asymptotically approach zero (Gruber and van der Laan 2010a). For ease of exposition we will refer to the negative log likelihood loss function in our ATE example, $P_n(\mathscr{L}(\bar{Q}))$, where

$$\mathscr{L}(\bar{Q}) = -\{Y \log[1 - \bar{Q}(A, W)] + (1 - Y) \log[1 - \bar{Q}(A, W)]\}.$$

The first targeted forward selection step is to construct a model for g that conditions on a single covariate. There are K possibilities, each of which improves the likelihood for g compared to an intercept-only model. From these we select the covariate that most improves the goodness of fit for Q. We do this by carrying out the TMLE updating step for all K univariate models for g, and evaluating the empirical loss for each, $\mathscr{L}(\bar{Q}^*_{n,k})$. Identify k^* as the minimizer of this loss. The first model for g in our sequence has now been identified, and the corresponding first candidate TMLE in the sequence has been defined.

The second model for g is constructed in a similar fashion. Now we consider all $J = K - 1$ bivariate models for g that have W_{k^*} included by default. In conjunction with $\bar{Q}^0_n$, these models index tentative candidate TMLEs $\bar{Q}^*_{n,j}$, where this time j runs from 1 to $K - 1$. As before, we evaluate $\mathscr{L}(\bar{Q}^*_{n,j})$ for each tentative candidate and find j^* that minimizes the loss. Next we compare $\mathscr{L}(\bar{Q}^*_{n,j^*})$ with $\mathscr{L}(\bar{Q}^*_{n,k^*})$. If $\mathscr{L}(\bar{Q}^*_{n,j^*}) \leq \mathscr{L}(\bar{Q}^*_{n,k^*})$ this forward selection step is complete. However, if $\mathscr{L}(\bar{Q}^*_{n,j^*}) > \mathscr{L}(\bar{Q}^*_{n,k^*})$ we conclude that no bivariate model for g offers an improvement over a univariate model for g.

If we were restricted to using a single fluctuation covariate to update $\bar{Q}^0_n$, the forward selection procedure would have to end. However, allowing a second fluctuation guarantees improving the likelihood for Q, and may also improve estimation of ψ. This is accomplished by fixing $\varepsilon_{n,1}$ at its previously estimated value, and adding a new fluctuation covariate to the model for $\bar{Q}$, $\bar{Q}^*_n(A, W) = \bar{Q}^0_n(A, W) + \varepsilon_{n,1} H_{g_{n_{k^*}}}(A, W) + \varepsilon_2 H_{g_{n_{j^*}}}(A, W)$ (on the logit scale). ε_2 is fit by maximum likelihood, ensuring that the efficient influence curve equation is solved at this new g. Another way to think of this is that at step j we update our notion of the baseline $\bar{Q}^0_n$ to include the fixed prior fluctuation. Let $\tilde{Q}_{n,k^*}(A, W) = \text{expit}\{\text{logit}[\bar{Q}^0_n(A, W)] + \varepsilon_{n,1} H_{g_{n_{k^*}}}(A, W)\}$, and we have that $\bar{Q}^*_n(A, W) = \text{expit}\{\text{logit}[\tilde{Q}_{n,k^*}(A, W)] + \varepsilon_2 H_{g_{n_{j^*}}}(A, W)\}$.

Targeted forward selection proceeds in this manner until all K covariates have been incorporated into the model for g. Then V-fold cross validation is used to select the best candidate in the corresponding sequence of TMLEs. The data are partitioned into V folds of approximately equal size, n/V. Each fold v is held out as a validation set, $val(v)$, in turn, with the remaining observations constituting the training set. Targeted forward selection algorithm is run on the training set, with $\bar{Q}^*_{n,k}$ evaluated for observations in validation set $val(v)$. The C-TMLE corresponds to the candidate TMLE in the sequence that minimizes the cross-validated loss, such as the cross-validated negative log likelihood,

$$L_{cv}(\bar{Q}) = -\frac{1}{n} \sum_{v=1}^{V} \sum_{i \in val(v)} \left\{ Y_i \log \bar{Q}_{n,v}(A_i, W_i) + (1 - Y_i) \log[1 - \bar{Q}_{n,v}(A_i, W_i)] \right\},$$

where $\bar{Q}_{n,v}(A, W)$ is a fit based on observations in training set v.

The greedy approach outlined in Algorithm 1 satisfies the core elements of a C-TMLE. The sequence of nested g estimators satisfies the requirement that

Algorithm 1 Greedy C-TMLE

procedure GREEDY C- TMLE
Stage 1. Obtain initial estimate $\bar{Q}_n^0$
Stage 2. TARGETED FORWARD SELECTION to create a sequence of g estimators $\{g_{n,1}, \ldots, g_{n,K}\}$ and candidate TMLEs, $(\bar{Q}_{n,1}^*, \ldots, \bar{Q}_{n,K}^*)$
Find minimizer of cross validated loss: $k^* = \underset{k}{\text{argmin}}\ \mathscr{L}_{cv}(\bar{Q}_{n,k}^*)$
Evaluate the parameter estimate: $\psi_n = \Psi(\bar{Q}_{n,k^*}^*)$.
end procedure

procedure TARGETED FORWARD SELECTION
Initializations: $\tilde{Q}_n = \bar{Q}_n^0,\ \mathbf{W}' = \emptyset,\ g_n = P(A = 1),\ k = 1$
while $k \leq K$ **do**
for all $W_j \in \mathbf{W}$ **do**
Calculate $H_{g_{n,w}}$, where $g_{n,j}(A, \tilde{\mathbf{W}}, W_j) = P(A = 1 \mid \mathbf{W}', W_j)$
Evaluate targeted $\bar{Q}_{n,j}^* = \text{expit}[\text{logit}(\tilde{Q}_n) + \varepsilon_{n,j} H_{g_{n,j}}]$
end for
Find minimizer of empirical loss: $j^* = \underset{j}{\text{argmin}}\ \mathscr{L}(\bar{Q}_{n,j}^*)$
Set $\bar{Q}_{n,k}^* = \bar{Q}_{n,j^*}^*$
if $\mathscr{L}(\bar{Q}_{n,k}^*) > \mathscr{L}(\bar{Q}_{n,k-1}^*)$ **then**
Update baseline: $\tilde{Q}_n = \text{expit}[\text{logit}(\tilde{Q}_n) + \varepsilon_{n,k-1} H_{g_{n,k-1}}]$
else
$\mathbf{W} = \mathbf{W} - W_{j^*}$
$\mathbf{W}' = \mathbf{W}' + W_{j^*}$
$k = k + 1$
end if
end while
end procedure

$\{g_n^1, \ldots, g_n^K\}$ grows towards and arrives at a consistent estimator of g_0. Construction of the sequence of g estimators ensures g_n^{k+1} is a better empirical fit for g than g_n^k. Each forward selection step maximizes (over all possible moves) the increase in fit over the TMLE update step relative to its initial estimator, while also improving the fit for g. Terms are incorporated into the model for g for a single fluctuation covariate until there is a decrease in the likelihood for Q. At that point the sub-model is extended to include an additional covariate. Earlier fluctuation parameter values are fixed, and a new fluctuation parameter is fit by maximum likelihood.

The targeted forward selection algorithm chooses the strongest confounder first. Covariates are re-ordered at each subsequent step with respect to their impact on the shrinking residual bias. This re-ordering delays incorporation of a covariate highly correlated with one that is already in the model for g, and covariates that are not associated with the outcome (instrumental variables). Targeted forward selection results in a sequence of candidate TMLEs. Cross-validated loss $\mathscr{L}_{cv}(\bar{Q}_{n,k}^*)$ is evaluated based on out of sample predictions for all n observations for each k from 1 to K. Let k^* index the candidate minimizing $\mathscr{L}_{cv}(\bar{Q}_{n,k}^*)$. The corresponding candidate TMLE fit on all data is plugged in to the mapping to evaluate $\psi_n = \Psi(\bar{Q}_{n,k^*}^*)$.

1.4.1.1 Implementation Notes

The model for g is not restricted to main terms only, but may include higher-order interactions and transforms of the data. If time considerations preclude defining a rich SL library to estimate g at each step there are alternatives. One is to use a main terms logistic regression model. Another is to define an SL library that includes logistic regression models that regress treatment on different sets of covariates that include transforms, such as splitting categorical or continuous covariates into multiple binary covariates, and defining higher order terms.

A complementary approach is to create an augmented covariate set, W_{aug}, that in addition to W includes select interaction terms, and perhaps a matrix of externally-estimated SL fits for g bounded away from (0, 1) at different levels. Running the C-TMLE algorithm on W_{aug} allows more non-parametric modeling of W under a main terms logistic regression modeling paradigm, and has been shown to be helpful in some simulated scenarios (Sekhon et al. 2011).

The number of fluctuation parameters depends on characteristics of the data and cannot be known in advance. A minor variant of this greedy C-TMLE algorithm is to select the ideal number of steps, k^*, as described above, then evaluate the final C-TMLE using a single update to the initial $\bar{Q}_n^0$ such that $\bar{Q}_n^*$ is set to expit[logit$(\bar{Q}_n^0)$ + $\varepsilon_n H_{g_{k^*}}$]. For both variants the linear span of the score equations generated when fitting ε includes an efficient influence function based on the largest selected model for g. Simulation studies to date have not demonstrated compelling performance differences.

1.4.1.2 Simulation Study

Porter, 2011 illustrated how issues of misspecification bias, near positivity violations, and lack of boundedness on the problem are addressed by TMLEs and other DR estimators proposed in the literature (Porter et al. 2011). That article applied many recently developed DR estimators to estimate a population mean outcome under missingness using a Monte Carlo simulation study design proposed earlier in the literature (Kang and Schafer 2007a). The outcome and missingness probabilities were simple functions of independent and identically distributed (i.i.d.) normally distributed covariates $Z = (Z_1, Z_2, Z_3, Z_4)$. However, $n = 1000$ observations in the analytic dataset were given by $O = (\Delta Y, \Delta, W)$, where covariates $W = (W_1, W_2, W_3, W_4)$ were complex non-linear functions of the actual covariates Z.

DR estimators investigated include weighted least squares (WLS), augmented IPCW (A-IPCW), (Robins and Zhao 1994) bounded Horvitz Thompson (BHT), (Robins et al. 2007) a parametric regression based estimator (PRC) (Scharfstein et al. 1999; Robins 1999), a DR estimator that internally enforces bounds on g (Cao), (Cao et al. 2009) a bounded DR estimator developed by Tan that incorporates either a weighted least square outcome regression model (TanWLS) or the empirical efficiency estimator of van der Laan and Rubin (TanRV) (Tan 2006, 2010; Rubin

and van der Laan 2008), and TMLE and C-TMLE, with and without super learning to mitigate model misspecification.

Box plots of bias in 250 Monte Carlo estimates illustrate the relative performance of these estimators under misspecification of Q and correct specification of g, and also under dual misspecification of Q and g (Fig. 1.2). For both analyses $g_n(1, W)$ was bounded away from 0 at level 0.025. The unweighted main term linear regression model regressing Y on W produces a biased OLS estimate due to model misspecification and informative missingness. Misspecification bias is exacerbated by weighting with inverse probability weights based on values of $g_n(1, W)$ that are near 0 (Porter et al. 2011, adapted from their Fig. 2).

In contrast to OLS, all DR estimators are unbiased when g is correctly specified. Variation in the spread around the median estimate is largest for PRC and smallest for the Tan estimators. Under dual misspecification most of the DR estimators are more biased than OLS, with Cao, TanRV, and C-TMLE, exhibiting less bias than the others.

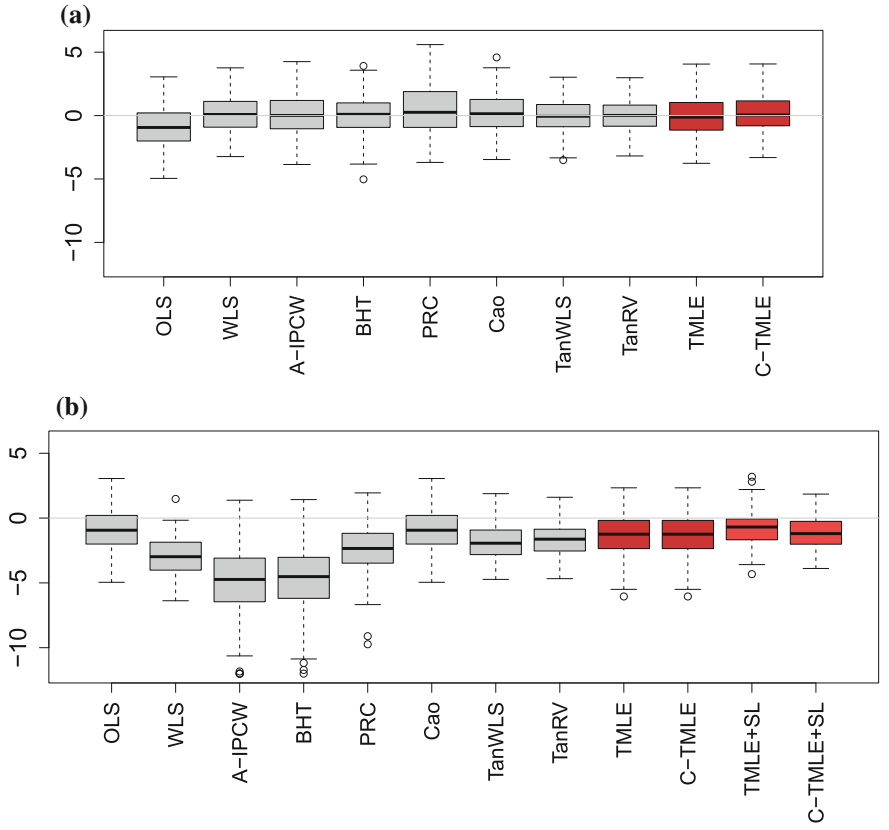


Fig. 1.2 Distribution of bias under misspecification of Q and correct specification of g (**a**), and under dual misspecification (**b**)

Algorithm 2 Scalable C-TMLE

Stage 1. Obtain initial estimate $\bar{Q}_n^0$
Stage 2.
Initialize $\tilde{Q}_n = \bar{Q}_n^0$, $k = 1$
Impose an ordering over covariates $W_1, \ldots, W_K$
procedure CONSTRUCT SEQUENCE OF CANDIDATE TMLES
  while $k \leq K$ **do**
    Estimate $g_{n,k}(A, W) = P(A = 1 \mid W_1, \ldots, W_k)$
    Evaluate targeted $\bar{Q}_{n,k}^* = \text{expit}[\text{logit}(\tilde{Q}_n) + \varepsilon_{n,k} H_{g_{n,k}}]$
    if $\mathscr{L}(\bar{Q}_{n,k}^*) > \mathscr{L}(\bar{Q}_{n,k-1}^*)$ **then**
      $\tilde{Q}_n = \text{expit}[\text{logit}(\tilde{Q}_n) + \varepsilon_{n,k-1} H_{g_{n,k-1}}]$
      Evaluate at new baseline: $\bar{Q}_{n,k}^* = \text{expit}[\text{logit}(\tilde{Q}_n) + \varepsilon'_{n,k} H_{g_{n,k}}]$
    end if
    $k = k + 1$
  end while
end procedure
Identify the best candidate TMLE: Find minimizer of cross validated loss: $k^* = \underset{k}{\text{argmin}}\ \mathscr{L}_{cv}(\bar{Q}_{n,k}^*)$
Evaluate the parameter estimate: $\psi_n = \Psi(\bar{Q}_{n,k^*}^*)$.

A strength of TMLE is its ability to incorporate machine learning. The two box plots on the far left of Fig. 1.2b illustrate that combining SL with TMLE and C-TMLE further reduced bias while also improving efficiency. C-TMLE+SL traded off bias and variance differently than TMLE+SL. Median bias of TMLE+SL was closer to 0, but three problematic datasets with outlying results extending beyond the whiskers on the TMLE+SL box plot were analyzed more successfully by C-TMLE+SL.

1.4.2 Scalable C-TMLE

The greedy C-TMLE algorithm described in the previous section does not scale well to high dimensional data. If there are k covariates, the algorithm's time complexity is $\mathscr{O}(k^2)$. Ju and colleagues proposed several alternate C-TMLEs with time complexity $\mathscr{O}(k)$ (Ju et al. 2016a, b). The common characteristic of these scalable C-TMLE algorithms is that they pursue a less aggressive strategy for ordering the sequence of g estimators. At each step k the greedy targeted forward selection strategy evaluates the bias/variance trade-off for ψ with respect to multivariate relationships among all $k - 1$ covariates previously incorporated into the model for g, treatment and the outcome. In contrast, scalable C-TMLEs impose a one-time pre-ordering of covariates at the outset. The decision at step k is merely whether or not an additional fluctuation covariate is warranted.

Given a method for imposing an ordering over covariates covariates, the basic algorithm for a scalable C-TMLE follows the procedure outlined in Algorithm 2.

We can verify that this general scalable C-TMLE algorithm meets the guidelines previously laid out in Sect. 1.3.2. First note that this algorithm constructs a sequence of nested g estimators, which necessarily increases in likelihood for g. These estimators non-parametrically grow towards the true g_0, and flexible modeling using SL or the more limited techniques mentioned earlier helps meet that requirement in practice. The final requirement that the the sequence of updated candidate TMLEs, $(\bar{Q}^*_{n,1}, \ldots, \bar{Q}^*_{n,K})$ is increasing in likelihood for Q, is met in Step 3c.

Unlike the greedy C-TMLE, the scalable C-TMLE imposes a pre-ordering on covariates that only takes advantage of collaborative double robustness at the outset. This pre-ordering will account for univariate impact on residual confounding, but ignores multivariate effects on estimator bias and variance. There are many possible pre-ordering schemes. For example, the first of two scalable C-TMLEs developed by *Ju, et. al.* pre-ordered covariates based on the log likelihood loss (Ju et al. 2016a). This is identical to the loss function evaluated in the first targeted forward selection step of the greedy C-TMLE algorithm. The loss function is evaluated for all K univariate models for g. This loss is used to imposes an ordering over covariates that ensures $L(\bar{Q}_{n,k})^* \leq L(\bar{Q}^*_{n,k+1})$, for $1 \leq k \leq K$.

Their second pre-ordering scheme ranks covariates according to the absolute value of the partial correlation $\rho_{YW_k \cdot A}$ between each covariate and the outcome within strata defined by treatment.

$$\rho(YW_k \cdot A) = \frac{\rho(R, W_k) - \rho(R, A) \times \rho(W_k, A)}{\sqrt{(1 - \rho(R, A)^2)(1 - \rho(W_k, A)^2)}}$$

where ρ is the Pearson correlation coefficient and $R = Y - \bar{Q}^0_n(A, W)$.

When all suspected confounders in W have a binary representation Bross's multiplicative bias formula provides an estimate of the strength of confounding. Covariates are ordered with respect to the absolute value of the log of multiplicative bias B_k (Bross 1954).

$$B_k = \frac{P_{W_k 1}(RR_{W_k Y} - 1) + 1}{P_{W_k 0}(RR_{W_k Y} - 1) + 1},$$

where $P_{W_k 1}$ is the mean value of suspected binary confounder W_k within the treated population, and $P_{W_k 0}$ is the mean value of W_k within the comparator population. $RR_{W_k Y}$ is the crude relative risk of the outcome associated with W_k, or optionally a relative risk estimate adjusted for other covariates. Covariates that have a large value for the Bross multiplier are considered the most important confounders and positioned earlier in the ordering.

1.4.2.1 Comparison of TMLE and C-TMLE Performance

The greedy and scalable C-TMLE procedures may not produce the same covariate ordering. Thus they may exhibit different finite sample performance. An SL-

Table 1.1 Monte Carlo simulation study results from Ju et al. under misspecification of Q and correctly specified g

	Bias	SD	MSE
TMLE	1.31	1.21	3.17
Greedy C-TMLE	0.25	1.01	1.27
LogLik C-TMLE	0.36	0.88	0.90
PartCor C-TMLE	0.32	0.92	0.95
SL-C-TMLE	0.37	0.88	0.90

based C-TMLE (SL-C-TMLE) was defined that uses cross-validation to identify the C-TMLE that performs best on a given dataset (Ju et al. 2016b). The SL-CTMLE is scalable (computational complexity $\mathcal{O}(k)$) when all candidate C-TMLEs under consideration are scalable. Ju and colleagues compared the performance of TMLE, and several C-TMLEs in analyses of simulated and real world data (Ju et al. 2016b). In many scenarios the various C-TMLEs performed equally well. Performance differences were apparent in Simulation Study 4, where there were near positivity violations. In this study six i.i.d, covariates $W = (W_1, \ldots, W_6) \sim N(0, 1)$ were made available to each estimator. Treatment probabilities were generated as $g_0(1, W) = \text{expit}(2W_1 + 0.2W_2 + 3W_3)$, and continuous Y was generated as $Y = A + 0.5W_1 - 8W_2 + 9W_3 - 2W_5 + N(0, 1)$.

Only three of the six covariates are true confounders, (W_1, W_2, W_3). W_4 and W_6 are unrelated to Y and A, and W_5 is predictive of Y. The initial regression model for Q was specified as a regression of Y on A, W_1, W_2. Residual bias due to model misspecification and lack of adjustment for all confounding remains a function of W_3, W_1, and W_2. Results of a Monte Carlo simulation study (1000 replicates, n = 1000) show that the C-TMLEs were able to reduce finite sample bias and variance relative to TMLE using the true g_0 (Table 1.1). Each C-TMLE made a slightly different bias/variance tradeoff. The greedy C-TMLE was least biased, while overall MSE was minimized by the scalable SL-C-TMLE and the scalable C-TMLE using a log likelihood pre-ordering scheme.

1.5 Applications of C-TMLE in Health Care

1.5.1 Biomarker Discovery

We previously described an application of C-TMLE to assess genomic variable importance (Gruber and van der Laan 2010a). An antiretroviral drug may be more effective in suppressing human immunodeficiency virus (HIV) in some strains of HIV than others. C-TMLE was applied to the problem of identifying mutations in viral DNA that might affect the virus's response to lopinavir. The data consisted of

Table 1.2 Mutation name, gold standard Stanford score (2007), C-TMLE estimate for each mutation. C-TMLE classification as likely/unlikely to confer resistance agrees with the Stanford score for mutations shown in blue

Mutation	Score	Estimate	Mutation	Score	Estimate
p50V	20	1.70*	p53LY	3	0.21
p82AFST	20	0.39*	p73CSTA	2	0.64*
p54VA	11	0.51*	p24IF	2	0.23
p54LMST	11	0.37*	p10FIRVY	2	−0.27
p84AV	11	0.10	p71TVI	2	0.02
p46ILV	11	0.05	p23I	0	0.82
p82MLC	10	1.61*	p36ILVTA	0	0.27
p47V	10	0.81*	p16E	0	0.24
p84C	10	0.60*	p20IMRTVL	0	0.18
p32I	10	0.54*	p63P	0	−0.13
p48VM	10	0.31	p88DTG	0	−0.43*
p90M	10	0.21	p30N	0	−0.44*
p33F	5	0.30	p88S	0	−0.47*

*95% CI excludes the null

401 observations $O = (Y, W)$, where outcome Y is the change in $\log_{10}$ viral load between baseline and post-treatment followup, and W contains binary indicators for the presence of 26 separate mutations, and an additional 51 baseline characteristics and treatment history covariates. 26 separate analyses were carried out, where exposure A was set to each one of the mutation indicators in turn, and the other 25 mutation indicators were included in the adjustment set. The goal was to assess the impact of each mutation on the change in HIV viral load (Bembom et al. 2008).

Positivity violations due to high correlations among mutations and a rare mutations make it impossible to estimate a true causal effect. Nevertheless, variable importance association measures (VIM) would help rank mutations according to their impact on the outcome. C-TMLE was the right tool to address the challenge of identifying an adjustment set that would provide a stable, low variance estimates of the association between A and Y. Results were compared with a 2007 gold standard assessment known as the Stanford score (0–20, with 20 indicating highly associated with

resistance as of September, 2007, subsequently modified. http://hivdb.stanford.edu). Table 1.2 shows C-TMLE VIM estimates for each mutation, ranked according to the Stanford score. Starred estimates were statistically significant at level $\alpha = 0.05$. C-TMLE was able to correctly classify most mutations as being likely or unlikely to confer resistance to lopinavir. C-TMLE found statistically significant positive effects for 8 of 12 mutations with Stanford score of 10 or above. C-TMLE analyses were consistent with the null hypothesis of no effect or a protective effect on resistance for 12 of 13 mutations with Stanford scores below 5.

1.5.2 Drug Safety

Clinical studies required for drug approval by a regulatory agency such as the FDA are typically under-powered for detecting rare adverse events and small risk increases (Singh and Loke 2012). Large post-licensure observational studies fill an important gap in conducting drug safety surveillance. C-TMLE was used to assess the impact of pioglitazone versus sulfonylurea on acute myocardial infarction (AMI) in a new user population of diabetic patients without prior cardiovascular disease (Lendle et al. 2013). The outcome of interest was the occurrence of AMI within six months of drug initiation. Minimal loss to follow-up was deemed ignorable, and C-TMLE was used to analyze a complete case dataset.

Data on $n = 27,168$ patients seen at Kaiser Permanente Northern California consisted of observations $O = (Y, A, W)$. Outcome Y was set to 1 for patients who experienced an AMI within 6 months, and 0 for those who did not. A was a binary indicator of treatment with pioglitazone ($A = 1$, $N_1 = 2,146$) versus comparator sulfonylurea ($A = 0, N_0 = 25,022$). Covariate vector W consisted of approximately 50 covariates including demographic information, comorbidities, and other drugs identified by experts as potential confounders. The outcome was rare, with only 5 (0.23%) occurrences in the treatment group, and 85 (0.34%) in the comparator group.

Table 1.3 Additive treatment effect of pioglitazone versus sulfonylurea on acute myocardial infarction in a new user cohort of diabetic patients with no prior cardiovascular history

Method	Estimate	Standard Error	P-value
Unadjusted	−0.0011	0.0013	0.39
Outcome regression	−0.0007	0.0014	0.61
PSM	−0.0013	0.0017	0.45
IPTW	−0.00005	0.0015	0.75
AIPTW	−0.0003	0.0015	0.86
TMLE	−0.0004	0.0015	0.80
C-TMLE	−0.0010	0.0011	0.38

Abbreviations: *PSM* propensity score matching, *IPTW* inverse probability of treatment weighting, *AIPTW* augmented-IPTW, *TMLE* targeted minimum loss-based estimation, *C-TMLE* collaborative TMLE

The unadjusted ATE estimate was −0.0011 (Table 1.3, reproduced from Lendle et al. (2013) (their Table 2). Adjusting for covariates in an outcome regression model moved the point estimate further away from the null, with a similar standard error. Propensity score-based estimators and DR estimators were closer to the null. Without knowing the truth bias cannot be assessed, however, C-TMLE had the smallest standard error. No point estimates were statistically significant. Results support a substantive conclusion of no evidence of difference in risk between the two drugs studied.

1.5.3 Future Work

1.5.3.1 C-TMLE for Multiple Time Point Interventions

In a point treatment setting with non-ignorable missingness in the outcome the data structure is given by $O = (\Delta Y, \Delta, A, W) \sim P_0$. Δ is a binary indicator of whether the outcome is observed. $\Delta Y = Y$ when Δ is 1, and is missing when $\Delta = 0$. In this data structure g factorizes as (g_A, g_Δ). g_A is the component of g discussed up until now. g_Δ is the distribution of missingness conditional on A and W. The targeting step for a TMLE to estimate ATE in this data structure uses an updated fluctuation covariate $H'_g(A, W)$, that is a function of both components of g. $H'_g(A, W) = \Delta / P(\Delta = 1 \mid A, W) H_g(A, W)$, with $H_g(A, W)$ as defined in Sect. 1.2. When (A, W) is not highly predictive of missingness, the same straightforward extension can be applied in the C-TMLE targeting step. Missingness probabilities are estimated externally and incorporated into $H'_{g_{n,k}}(A, W)$ constructed by a greedy or scalable C-TMLE.

This simple example of factorizing g generalizes to longitudinal data analysis, with multiple opportunities for treatment decisions and loss to follow-up. TMLEs to estimate the effects of multiple time point interventions have been developed (van der Laan and Gruber 2012; Schnitzer et al. 2014; Petersen et al. 2014). An application of C-TMLE in survival analysis appears in the literature (Stitelman and van der Laan 2010), however because g factorizes into t components, where t is the total number of treatment and censoring opportunities, the approach is computationally intensive and does not scale. Extending scalable C-TMLEs to the analysis of longitudinal data is an active area of current research.

1.5.3.2 Further Robustifying C-TMLE

Theory suggests another fruitful avenue for C-TMLE. The estimating equation $P_0 D(Q_n^*, g_n^*, \psi_0) = 0$ can be equivalently expressed as

$$P_0 \left[D(Q_n^*, g_n^*, \psi_0) - D(Q_0, g_n^*, \psi_0) \right] = 0. \tag{1.3}$$

The efficient influence curve can be decomposed as $D(Q, g, \psi) = D_{IPTW}(g, \psi) - D_{CAR}(Q, g)$ (Theorem 1.3, van der Laan and Robins (2003); Robins and Rotnitzky (1992)). This decomposition allows rearrangement of Eq. (1.3) to give $P_0 D_{CAR}(Q_n^* - Q_0, g_n^*) = 0$. Due to linearity of $Q \to D_{CAR}(Q, g)$, Eq. (1.3) is equivalent with $P_0 D_{CAR}(Q_n^* - Q_0, g_n^*) = 0$ (Theorem 1, van der Laan and Gruber (2010)). In other words, to be effective C-TMLE should target solving $P_0 D_{CAR}(\bar{Q}_n - \bar{Q}_0, g_n) = 0$. For the ATE parameter $D_{CAR}(Q - Q_0, g) = H_g(\bar{Q} - \bar{Q}_0)(A - g(1 \mid W))$, with

$$H_g(\bar{Q} - \bar{Q}_0) \equiv \frac{(\bar{Q} - \bar{Q}_0)(1, W)}{g(1 \mid W(Q))} + \frac{(\bar{Q} - \bar{Q}_0)(0, W)}{g(0 \mid W(Q))}.$$

This representation of D_{CAR} teaches us how to construct a targeted estimator (g_n^*, Q_n^*) such that $P_n D^*(Q_n^* - Q_0, g_n^*) = 0$. For example, for the ATE parameter the joint TMLE of (g_n, Q_n^*) involves iteratively fitting a logistic regression of A given W with submodel $\text{logit}(g_n^{k,\varepsilon}) = \text{logit}\, g_{n.k} + \varepsilon H_{g_n,k}(Q_n^k - Q_0)$, simultaneously with the usual TMLE update of Q_n^k. We refer to this joint TMLE as an oracle C-TMLE since it uses an unknown (oracle) covariate that captures the residual bias $(Q_n - Q_0)$ required for consistent estimation of ψ_0. This oracle C-TMLE is always consistent, even when g_n and Q_n are both inconsistent. Starting with a targeted C-TMLE, information in $f_0(A, W) = \bar{Q}_n^0(A, W) - \bar{Q}_0(A, W)$ is used to simultaneously update Q and g. This updating step is iterated until convergence. For example, the oracle updating procedure for the ATE target parameter is given by,

$$\begin{aligned}
\text{logit}(g_n^{m+1}) &= \text{logit}(g_n^m) + \varepsilon_1 H_{CAR}(f_0, g_n), \\
\text{logit}(\bar{Q}_n^{m+1}) &= \text{logit}(\bar{Q}_n^m) + \varepsilon_2 H_{g_n}^*(A, W),
\end{aligned}$$

where

$$H_{CAR}(f, g) = \frac{f(1, W)}{g(1 \mid W)} - \frac{f(0, W)}{g(0 \mid W)}, \text{ and } H_{g_n}^*(A, W) = \frac{2A - 1}{g(A \mid W)}.$$

Of course, $\bar{Q}_0$ is not known, so this approach is not available for analyses of real world data. However, a realistic C-TMLE is obtained by estimating the unknown residual bias $(Q_n^k - Q_0)$.

We demonstrate the properties of an oracle version of this estimator that can evaluate the required value of the residual $(\bar{Q}_n - \bar{Q}_0)$. We used data generated in accordance with simulation study 4 from a paper by Van Steelandt and colleagues to evaluate estimator performance (Vansteelandt et al. 2012). That paper proposed a propensity score estimation procedure incorporating $f(A, W)$ that guarantees consistency under dual misspecification of Q and g when $\bar{Q}_0$ happens to be a linear combination of $f(A, W)$ components. This regression doubly robust (RDR) IPW estimator is exploiting collaborative double robustness. Results of a simulation study comparing performance of a crude and adjusted OLS estimator, G-estimation (G) (Robins et al. 1992), IPW, RDR, augmented IPW (A-IPW), and the greedy C-TMLE

Table 1.4 Results of simulation study under dual misspecification of Q and g not using $f_0(A, W)$ (left) and incorporating information in the true $f_0(A, W)$ (right).

$\bar{Q}_n$ misspecified							
	g_n misspecified				Using true $f_0(A, W)$		
	Bias	Var	MSE		Bias	Var	MSE
OLS (unadj)	0.99	0.05	1.03				
OLS (adj)	0.31	0.03	0.12				
G	−0.07	0.01	0.02				
IPW	−0.08	1.11	1.11	IPW_s	−0.17	0.03	0.06
RDR	0.02	0.10	0.10	RDR_s	0.21	0.02	0.06
AIPW	−0.29	17.33	17.40	AIPW_s	0.15	0.02	0.04
C-TMLE_{greedy}	−0.12	0.02	0.04	C-TMLE_{oracle}	0.02	0.01	0.01

are shown on the left hand side of Table 1.4. Results on the right hand side of the table are for novel versions of these estimators described in *Vansteelandt, et. al.* that make use of $f(A, W)$ to stabilize the propensity scores, and also the oracle C-TMLE_{oracle}. For all estimators $f(A, W)$ was estimated from data using correctly specified models. This greatly reduced MSE for all enhanced estimators, with C-TMLE_{oracle} having the smallest bias, variance, and MSE.

The insight that Q and g together can account for all bias due to confounding even when neither on its own can be used to consistently estimate ψ_0 is very powerful. The C-TMLEs presented in this chapter already avoid adjusting for covariates in the nuisance parameter model that will inflate finite sample bias and variance. This new insight that we do not even need to condition on $(Q_n - Q_0)$, but instead can just solve the right score equation, $P_0 D_{CAR}(Q - Q_0, g) = 0$, motivates a new class of C-TMLEs. These offer an opportunity to reduce residual bias without adjusting for additional covariates in the outcome or censoring mechanism models. Along with the existing data-adaptive C-TMLEs that exploit collaborative double robustness, these promising new estimators can be important tools for analysis of high dimensional and sparse data.

References

Bembom, O., Fessel, J. W., Shafer, R. W., & van der Laan, M. J. (2008). Data-adaptive selection of the adjustment set in variable importance estimation. Technical report, U.C. Berkeley Division of Biostatistics Working Paper 231. http://biostats.bepress.com/ucbbiostat/paper231.

Bross, I. D. J. (1954). Misclassification in 2 × 2 tables. *Biometrics*, *10*, 478–486.

Burns, P. B., Rohrich, R. J., & Chung, K. C. (2011). The levels of evidence and their role in evidence-based medicine. *Plastic and Reconstructive Surgery*, *128*, 305–310.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Cao, W., Tsiatis, A. A., & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, *96*, 723–734.

Davidian, M., & Lunceford, J. K. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*, 2937–2960.

Freedman, D. A., & Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review*, *32*, 392–409.

Gruber, S., & van der Laan, M. J. (2010a). A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics, 6*(1).

Gruber, S., & van der Laan, M. J. (2010b). An application of collaborative targeted maximum likelihood estimation in causal inference and genomics.*The International Journal of Biostatistics, 6*(1).

Gruber, S., & van der Laan, M. J. (2015). Consistent causal effect estimation under dual misspecification and implications for confounder selection procedures. *Statistical Methods in Medical Research*, *24*(6), 1003–1008.

Hernan, M. A., Brumback, B., & Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, *11*(5), 561–570.

Ju, C., Combs, M., Lendle, S. D., Franklin, J. M., Wyss, R., Schneeweiss, S., & van der Laan, M. J. (2016a, June). Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods. Technical report, Division of Biostatistics, University of California, Berkeley.

Ju, C., Gruber, S., Lendle, S. D., Franklin, J. M., Wyss, R., Schneeweiss, S., & van der Laan, M. J. (2016b). Scalable collaborative targeted learning for large scale and high-dimensional data. Technical report, Division of Biostatistics, University of California, Berkeley.

Kang, J., & Schafer, J. (2007a). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, *22*, 523–539.

Kang, J., & Schafer, J. (2007b). Rejoinder: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 574–580.

King, G., Lucas, C., & Nielsen, R. (2014). The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science*.

Lendle, S. D., Fireman, B., & van der Laan, M. J. (2013). Targeted maximum likelihood estimation in safety analysis. *Journal of Clinical Epidemiology*, *66*(8), S91–S98.

Pearl, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology*, *174*(11), 1223–1227.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., & van der Laan, M. J. (2010). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, published online 28 Oct. https://doi.org/10.1177/0962280210386207.

Petersen, M. L., Schwab, J., Gruber, S., Blaser, N., Schomaker, M., & van der Laan, M. J. (2014). Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models.*Journal of Causal Inference, 2*(2), 147–185.

Porter, K. E., Gruber, S., van der Laan, M. J., & Sekhon, J. S. (2011). The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics*, *7*(31), 1–34.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association: Section on Bayesian Statistical Science*, pp. 6–10.

Robins, J. M., & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology: Methodological Issues*. Bikhäuser.

Robins, J. M., Blevins, D., Ritter, G., & Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of aids patients. *Epidemiology*, 319–336.

Robins, J. M., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, *22*, 544–559.

Robins, J. M. (1999). Commentary on using inverse weighting and predictive inference to estimate the effecs of time-varying treatments on the discrete-time hazard. *Statistics in Medicine*, *21*, 1663–1680.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129.

Robins, J. M., & Rotnitzky, A. (2001). Comment on the Bickel and Kwon article, 'Inference for semiparametric models: Some questions and an answer'. *Statistica Sinica*, *11*(4), 920–936.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rotnitzky, A., Robins, J. M., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *The Journal of the American Statistical Association, 89*, 846–866.

Rubin, D. B., & van der Laan, M. J. (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics, 4*(1), Article 5.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for non-ignorable drop-out using semiparametric nonresponse models, (with discussion and rejoinder). *Journal of the American Statistical Association, 94*(448), 1096–1120 (1121–1146).

Schnitzer, M. E., van der Laan, M. J., Moodie, E. E. M., & Platt, R. W. (2014, 06). Effect of breastfeeding on gastrointestinal infection in infants: A targeted maximum likelihood approach for clustered longitudinal data. *Annals Applied Statistics, 8*(2), 703–725. https://doi.org/10.1214/14-AOAS727.

Sekhon, J. S. (2008). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Sotware* (Forthcoming).

Sekhon, J. S., Gruber, S., Porter, K., & van der Laan, M. J. (2011). Propensity-score-based estimators and C-TMLE. In van der Laan, M. J. & Rose, S. *Targeted learning: Prediction and causal inference for observational and experimental data*, chap. 21. New York: Springer.

Singh, S., & Loke, Y. K. (2012). Drug safety assessment in clinical trials: Methodological challenges and opportunities. *Trials, 13*.

Smith, J., & Todd, P. (2005). Does matching overcome lalondes critique of nonexperimental estimators? *Journal of Econometrics*, *125*, 305–353.

Stitelman, O. M., & van der Laan, M. J. (2010). Collaborative targeted maximum likelihood for time to event data. *The International Journal of Biostatistics, 6*(1).

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *The Journal of the American Statistical Association*, *101*, 1619–1637.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, *97*(3), 661–682.

van der Laan, M. J., & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. New York: Springer.

van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Prediction and causal inference for observational and experimental data*. New York: Springer.

van der Laan, M. J., & Gruber, S. (2010, January). Collaborative double robust penalized targeted maximum likelihood estimation. *The International Journal of Biostatistics, 6*(1).

van der Laan, M. J., & Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics, 8*.

van der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics, 2*(1).

van der Laan, M. J., Polley, E., & Hubbard, A. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology, 6*(25). ISSN 1.

Vansteelandt, S., Bekaert, M., & Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, *21*, 7–30.

Chapter 2
Generalized Tests in Clinical Trials

Stephan Ogenstad

2.1 Introduction

Conventional statistical methods do not provide exact solutions to many statistical problems, such as those arising in ANOVA, mixed models and multivariate analysis of variance (MANOVA), especially when the problem involves a number of nuisance parameters. As a result, users of these methods often resort to approximate or asymptotic statistical methods that are valid only when the sample size is large. With small or ordinary sample sizes, such methods often have poor performance (Weerahandi 1994). The approximate and asymptotic methods may lead to misleading conclusions or may fail to detect truly significance results from clinical studies.

Classical statistical tests may be insensitive to a wide range of situations occurring commonly in practice, particularly when the effect of the factor under study is heterogeneous. All statistical procedures are based on some distributional assumptions. In addition, many statistical procedures (e.g. ANOVA, ANCOVA) use the F-test and are based on the assumption of homoscedasticity (equal variances) and relate to the validity of the often convenient assumption that the structure of any one part of a dataset is the same as any other part. From experience, this assumption is seldom true when responses are different in the separate treatment groups. The assumption of equal variances is usually made for simplicity and mathematical ease rather than anything else. The outcome of using conventional statistical models when the assumptions are not reasonable can lead to serious consequences. In many situations, these procedures can fail to detect significant therapeutic effects even when available data provide sufficient evidence that the effects are present. In other applications, the conventional statistical models sometimes lead to incorrect conclusions, implying that the therapeutic results are significant when they are actually not (Blair and Higgins 1980; Brownie et al. 1990; Graubard and Korn 1987).

S. Ogenstad (✉)
Statogen Consulting LLC, 1600 Woodfield Creek Drive #215, Wake Forest, NC 27587, USA
e-mail: sogenstad@statogen.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_2

For instance, in the classical handling of the statistical problem in one-way ANOVA, it is assumed that the population variances are all equal. This is not really a natural assumption. In fact, it is often seen in most applications that the variances tend to be substantially different especially when the mean responses are substantially different. From simulation studies, it has also been observed that the assumption of equal variances is much more serious than the assumption of normally distributed populations, in that the former has the greater chance of leading to wrong conclusions. The classical ANOVA problems that rely on the equal variances assumption can dramatically reduce the power of the tests. Moreover, the magnitude of the lack of power problem of the tests based on the equal variance assumption increases with the number of treatments being compared. We also want to point out that in most applications, despite a common belief, it is not possible to transform data to achieve the approximate normality and equal variances simultaneously. The *p*-value produced from the classical approach is valid only if the variances are equal, and the test is not appropriate if the variances are significantly different.

In the analysis of repeated measures, it is also, assumed that all treatment groups have equal variances. While there is no serious problem when the assumption is reasonable, the assumption can lead to serious erroneous conclusions when the variances are substantially different. Moreover, in situations of higher-way ANOVA under an incorrect heteroscedasticity assumption, one is more prone to draw misleading conclusions. For instance, one can be misled by the classical F-test to conclude that a certain factor of an ANOVA is significant when in reality a different factor is significant.

Extensions have been made to the classical methods in repeated measures involving mixed models, MANOVA, and growth curves, in particular. Repeated measures and growth curves models are in fact special classes of mixed models. The classical approach to solving these problems provides exact solutions to only a fraction of the problems. Conventional methods alone do not always provide exact solutions to even some simple problems. For instance, in the univariate analysis of variance, the classical approach fails to provide exact tests when the underlying population variances are unequal. In some widely used growth curve models, there are no exact classical tests even in the case of equal variances. As a result, users of these methods often resort to asymptotic results in search of approximate solutions even when such approximations are known to perform rather poorly with moderate sample sizes.

Solutions to the statistical problems are addressed as extensions, as opposed to alternatives, to conventional methods of statistical inference. In Weerahandi (1994), each class of problems is started with a simple model under special assumptions that are necessary for the classical approach to work. After discussing solutions available for such special cases, these assumptions are relaxed when they are considered to be too restrictive or unreasonable in some applications, especially when they are known to have poor size (Type I error) or power performance. For instance, in fixed effects ANOVA, the problem is first considered under the homoscedastic variance/covariance assumption and then later the assumption is dropped.

The generalized methods are exact in the sense that the tests and the confidence intervals are based on exact probability statements rather than on asymptotic approx-

imations. This means that inferences based on them can be made with any preferred accuracy, provided that assumed parametric model or other assumptions are correct. To make this possible, solutions to problems of testing various hypotheses are presented in terms of p-values. There is readily available computer software to implement these exact statistical methods. Exact p-values and confidence intervals obtained with extended definitions also serve to provide excellent approximate solutions in the classical sense. From simulation studies reported in the literature, type I error and power performance of these approximations are usually much better than the performance of more complicated approximate tests obtained by other means.

By exact generalized inference, we mean various procedures of hypothesis testing and confidence intervals that are based on exact probability statements. Weerahandi (1994) uses the term '*exact*' rather than '*generalized*' methods because these methods are not approximations to the problems but exact solutions. Here we confine our attention to the problems of making inferences concerning parametric linear models with normally distributed error terms. In particular, we do not address exact nonparametric methods that are discussed, for instance in Good (1994) and Weerahandi (1994). The purpose of this chapter is to provide a brief introduction to the notions and methods in the generalized inference that enable one to obtain parametric analytical methods that are based on exact probability statements.

There is a wide class of problems for which classical fixed-level tests based on sufficient statistics do not exist, and there are simple problems in which conventional fixed-level tests do not exist. For instance, consider the mean μ and variance σ^2 in a normal distribution $N(\mu, \sigma^2)$ and let us assume that the parameter of interest is the second moment of the normal random variable X about a point other than the mean, say k, then the parameter of interest is

$$E(X-k)^2 = \mu^2 + \sigma^2 - 2k\mu + k^2.$$

Classical tests are not available for this parameter unless $k = \mu$ (Weerahandi 1994). If instead, the parameter of interest is $\theta = \mu + k\sigma^2$, then it is possible but not easy to find a test statistic whose value and distribution depends on the parameters only through the parameter of interest, since either μ or σ^2 can be considered as the nuisance parameter.

Actually, these kinds of problems are prevalent even with widely used linear models. For instance, in the problem of comparing the means of two or more normal populations, exact fixed-level tests and conventional confidence intervals based on sufficient statistics are available only when the population variances are equal or when some additional information is available about the variances. The situation only gets worse in more complicated problems such as the two-way ANOVA, the MANOVA, mixed models, and in repeated measures models including crossover designs and growth curves.

In the application of comparing two regression models, Weerahandi (1987) gave the first introduction to the notion of generalized p-value and showed that it is an exact probability of an unbiased extreme region, a well-defined subset of the sample space formed by sufficient statistics. Motivated by that application, Tsui and Weerahandi

(1989) provided formal definitions and methods of deriving generalized *p*-values. In a Bayesian treatment, Meng (1994) introduced a Bayesian *p*-value, as a posterior predictive *p*-value, which is, under the noninformative prior, numerically equivalent to the generalized *p*-value. Weerahandi and Tsui (1996) showed how Bayesian *p*-values could be obtained for ANOVA-type problems that are numerically equivalent to the generalized *p*-values.

As discussed in detail in Weerahandi (1994), exact probability statements are not necessarily related to the classical repeated sampling properties. In special cases, the former may have such implications on the latter, but this is not something that one should take for granted. For instance, in applications involving discrete distributions, often we can compute exact *p*-values, but not exact fixed-level tests. Rejecting a hypothesis based on such *p*-values, say at the 5% level if $p<0.05$, does not imply that the false positive rate in repeated sampling is 5%. Simply, such a *p*-value is a measure of false positive error and hence we can, in fact, reject the null hypothesis when it is less than a certain threshold. However, in most applications, fixed-level tests based on *p*-values, including the generalized *p*-values, do provide excellent approximate fixed-level tests that are better than asymptotic tests. Indeed, consistent with simulation studies reported in the literature (Gamage and Weerahandi 1998; Burdick et al. 2005), generalized tests based on exact probability statements tend to outperform, in terms of type I error or power, the more complicated approximate tests. Moreover, in many situations, type I error of generalized tests do not exceed the intended level. Therefore, procedures based on probability statements, that are exact for any sample size, are always useful, regardless of if we insist on repeated sampling properties or not. Also to those who insist on classical procedures, and anyone who has difficulties with the meaning of exactness, we can consider the generalized approach as a way of finding good approximate tests and confidence intervals, which are expected to perform better than asymptotic methods. We can benefit from the generalized approach to statistical inference, since it is an extension of the classical approach to inference as opposed to an alternative, providing solutions to a wider class of problems.

2.2 Test Variables and Generalized *p*-Values

Classical *p*-values as well as testing at a fixed nominal level, are based on what is known as test statistics. Basically, a test statistic is a function of some special properties of some observable dataset, that will distinguish the null from the alternative hypothesis. The function should not depend on any unknown parameters to qualify to be a test statistic. In the classical methodology of testing of hypotheses, this is an important requirement since, given a dataset, we should be able to compute such a statistic and compare against a critical value. Test statistics provide a convenient way of constructing extreme regions, on which *p*-values and tests can be based. But, this methodology only works in a very limited set of conditions (Weerahandi 1994). For instance, in the problem of sampling from a normal population, it is not clear how a

test statistic could be constructed if the parameter of interest were a function such as, $\theta = \mu + \sigma^2$. The Behrens-Fisher problem is a well-known example of a circumstance where a test statistic based on sufficient statistics does not exist when the variances are not assumed to be equal. This limitation extends well into all types of linear models including ANOVA, regression models, and all types of repeated measures problems.

Tsui and Weerahandi (1989) introduced the notion of *test variables* in the context of generalized inference. Test variables provide a convenient way of defining extreme regions as they play the role of test statistics in the generalized setting since test variables are extensions of test statistics.

A generalized p-value is an extension of the classical p-value, which except in a limited number of applications, provides only approximate solutions. Tests based on generalized p-values are exact statistical methods in that they are based on exact probability statements. While conventional statistical methods do not provide exact solutions to such problems as testing variance components or ANOVA under unequal variances, exact tests for such problems can be obtained based on generalized p-values (Gamage et al. 2013; Hamada and Weerahandi 2000; Krishnamoorthy et al. 2006). In order to overcome the shortcomings of the classical p-values, Tsui and Weerahandi (1989) extended the classical definition so that one can obtain exact solutions for such problems as the Behrens–Fisher problem and testing variance components. This is accomplished by allowing test variables to depend on observable random vectors as well as their observed values, as in the Bayesian treatment of the problem, but without having to treat constant parameters as random variables.

To provide formal definitions, consider a random vector $\mathbf{Y}$ with the cumulative distribution function $F(\mathbf{y}; \boldsymbol{\xi})$, where $\boldsymbol{\xi} = (\theta; \boldsymbol{\delta})$ is a vector of unknown parameters. θ is the parameter of interest and $\boldsymbol{\delta}$ is a vector of nuisance parameters. Let $\mathbf{y}$ be the observed value of the random vector $\mathbf{Y}$. An extreme region with the observed sample point on its boundary can be denoted as $C(\mathbf{y}; \theta, \boldsymbol{\delta})$. The boundary of extreme regions could be allowed to be any function of the quantities $\mathbf{y}, \theta$, and $\boldsymbol{\delta}$, and therefore, we need to allow test variables to depend on all these quantities. However, an extreme region is of practical use only if its probability does not depend on $\boldsymbol{\xi}$. Furthermore, a subset of the sample space obtained by more general methods should truly be an extreme region in that its probability should be greater under the alternative hypothesis than under the null hypothesis, as defined more formerly below.

Definition. A *generalized test variable* is a random variable of the form $T = T(\mathbf{Y}; \mathbf{y}, \boldsymbol{\xi})$ having the following three conditions:

1. The observed value $t = T(\mathbf{y}; \mathbf{y}, \boldsymbol{\xi})$ of T does not depend on unknown parameters.
2. The probability distribution of T does not depend on nuisance parameters.
3. Given t, $\mathbf{y}$ and $\boldsymbol{\delta}$, $P(T \leq t; \theta)$ is a monotonic function of θ.

2.3 Generalized Confidence Intervals

The classical approach to interval estimation suffers from more difficulties than that of hypothesis testing. Even when the problem does not involve nuisance parameters and there are exact confidence intervals, in some applications, they lead to results that contradict the very meaning of confidence. Both Ghosh (1961) and Pratt (1961) independently provided a very simple example of a uniformly most accurate confidence interval having highly undesirable properties, and connects two fundamental performance measures in confidence set estimation. Weerahandi (1994) showed how such undesirable confidence intervals can be avoided by expanding the class of intervals available to choose from. Just as in the case of testing of hypotheses, here we extend the class of available procedures for any given problem by insisting on exact probability statements rather than on sampling properties. This will enable us to solve such problems as the Behrens-Fisher problem for which exact classical confidence intervals do not exist. As in the Bayesian approach, the idea is to do the best with the observed data at hand instead of discussing other samples that could have been observed, was the process to be repeated. The generalized confidence intervals are nothing but the enhanced class of interval estimates obtained from exact probability statements with no special regard to repeated sampling properties that are of little practical use (Weerahandi 1994, 2004).

The definition of a confidence interval is generalized so that problems such as constructing exact confidence regions for the difference in two normal means can be undertaken without the supposition of equal variances. Under certain conditions, the extended definition is shown to preserve a repeated sampling property that a practitioner expects from exact confidence intervals. The proposed procedure can also be applied to the problem of constructing confidence intervals for the difference in two exponential means and for variance components in mixed models. With this description, we can carry out fixed level tests of parameters of continuous distributions on the basis of generalized p-values.

Thus, Weerahandi (1993) extended the conventional definition of a confidence interval in such a way that an applicably useful repeated sampling property is preserved. The research into this field was prompted by the need of exact confidence intervals in statistical problems involving nuisance parameters. For instance, even for a simple problem such as constructing confidence intervals for the difference in means of two exponential distributions, exact confidence intervals based on sufficient statistics are not available. The possibility of extending the definition of confidence intervals was suggested by the existence of p-values in this type of problem. Weerahandi (1987) used an extended p-value to compare two regressions with unequal error variances. The usefulness of generalized p-values explicitly defined by Tsui and Weerahandi (1989) is evident from a number of studies and applications, including those by Thursby (1992), Zhou and Mathew (1994), and Koschat and Weerahandi (1992).

To generalize the definition of confidence intervals, we first examine the properties of interval estimates obtained by the conventional definition. Consider a population

represented by an observable random variable Y. Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ be a random sample of n observations from the population. Suppose the distribution of the random variable Y is known except for a vector of parameters $\boldsymbol{\xi} = (\theta, \boldsymbol{\delta})$, where θ is a parameter of interest and $\boldsymbol{\delta}$ is a vector of nuisance parameters. We are interested in finding an interval estimate of θ based on observed values of $\mathbf{Y}$. The problem is to construct generalized confidence intervals of the form $[A(\mathbf{y}), B(\mathbf{y})] \subset \Theta$, where Θ is the parameter space and $A(\mathbf{y})$ and $B(\mathbf{y})$ are functions of $\boldsymbol{y}$, the observed data.

In the classical approach to interval estimation we find two functions of the observable random vector, say $A(\mathbf{Y})$ and $B(\mathbf{Y})$ such that the probability statement

$$P[A(\mathbf{Y}) \leq \theta \leq B(\mathbf{Y})] = \gamma, \tag{2.1}$$

is satisfied, where γ is specified by the desired confidence level.

If the observed values of the two statistics are $a = A(\mathbf{y})$ and $b = B(\mathbf{y})$, then $[a, b]$ is a confidence interval for θ with the confidence coefficient γ. For instance, if $\gamma = 0.95$, then the interval $[a, b]$ obtained in this manner is called a 95% confidence interval. If in the situation of interval estimation of the parameter θ, the interval could be constructed a large number of times to obtain new sets of observation vectors $\boldsymbol{y}$, then the confidence intervals obtained using the formula (2.1) will correctly include the true value of the parameter θ 95% of the times. After a large number of independent situations of setting 95% confidence intervals for certain parameters of interest, we will have correctly included the true value of the parameter in the corresponding intervals 95% of the times. It, of course, has no implication about the coverage based on the sample that we have actually observed. Indeed, Pratt (1961), Ghosh (1961), and Kiefer (1977) provide examples where the current intervals violating the very meaning of *confidence*. In particular, they showed that in those applications the so-called exact confidence intervals do not contain the parameters at all. The only thing truly exact about a confidence interval is the probability statement on which the interval is based. If indeed repeated samples can be obtained from the same experiment, then the claimed confidence level will no longer be valid and in the limit, the value of the parameter will be known exactly, so that statistical inference on the parameter is no longer an issue. In view of this, Weerahandi (1993) searched for intervals that would enhance the class of solutions and extended the class of candidates eligible to be interval estimators by insisting on the probability statement only. This will allow us to find interval estimates for situations where it is difficult or impossible to find A($\mathbf{Y}$) and B($\mathbf{Y}$) satisfying (1) for all possible values of the nuisance parameters. He further showed how this can be accomplished by making probability statements relative to the observed sample, as done in the Bayesian approach, but without having to treat unknown parameters as random variables. More precisely, we can allow $A()$ and $B()$ to depend on the observable random vector $\mathbf{Y}$ and the observed data $\mathbf{y}$ both. When there are a number of parameters of interest, in general, we could allow subsets of the sample space possibly depending on the current sample point $\mathbf{y}$ of $\mathbf{Y}$.

Such intervals Weerahandi referred to as *generalized confidence intervals*. The construction of such regions can be facilitated by generalizing the classical definition

of pivotal quantities. A random variable of the form $R = R(\mathbf{Y}; \mathbf{y}, \boldsymbol{\xi})$, a function of $\mathbf{Y}$, $\mathbf{y}$, and $\boldsymbol{\xi}$, is said to be a *generalized pivotal quantity* if it has the following two properties:

Property A: The probability distribution of R does not depend on unknown parameters.
Property B: The observed pivotal, $r_{obs} = R(\mathbf{y}; \mathbf{y}, \boldsymbol{\xi})$ does not depend on nuisance parameters $\boldsymbol{\delta}$.

Property A allows us to write probability statements leading to confidence regions that can be evaluated regardless of the values of the unknown parameters. Property B ensures that when we specify the region with the current sample point $\mathbf{y}$, then we can obtain a subset of the parameter space that can be computed without knowing the values of the nuisance parameters.

Suppose we have constructed a generalized pivotal $R = R(\mathbf{Y}; \mathbf{y}, \boldsymbol{\xi})$ for a parameter of interest and we wish to construct a confidence region at confidence coefficient γ. Consider a subset C_γ of the sample space chosen such that

$$P(R \in C_\gamma) = \gamma. \tag{2.2}$$

The region defined by (2.2) also specifies a subset $C(\mathbf{y}; \theta)$ of the original sample space satisfying the equation $P(\mathbf{Y} \in C(\mathbf{y}; \theta)) = \gamma$. Unlike classical confidence intervals, this region depends not only on γ and θ but also on the current sample point $\mathbf{y}$. With this generalization, we can obtain interval estimates on θ relative to the observed sample with no special regard to samples that could have been observed but were not. Although the generalized approach shares the same philosophy of the Bayesian approach that the inferences should be made with special regard to the data at hand, here we do not treat parameters as random variables and hence the probability statements are made with respect to the random vector $\mathbf{Y}$. Having specified a subset of the sample space relative to the current sample point, we can evaluate the region at the observed sample point and proceed to solve (2.2) for θ and obtain a region Θ_c of the parameter space that is said to be a $100\gamma\%$ generalized confidence interval for θ if it satisfies the equation

$$\Theta_c(r) = \{\theta \in \Theta | R(\mathbf{y}; \mathbf{y}, \xi) \in C_\gamma\},$$

where the subset C_γ of the sample space of R satisfies Eq. (2.2).

It should be reemphasized that generalized confidence intervals are not alternatives, but rather extensions of classical confidence intervals. In fact, for a given problem there is usually a class of confidence intervals satisfying the probability statement (2), a feature of classical intervals as well. Weerahandi (1994) discussed how the choice of appropriate generalized pivotals could be facilitated by invoking the principals of sufficiency and invariance. Even after we have obtained a particular pivotal quantity we could construct a variety of confidence regions. Depending on the application, a left-sided interval, a right-sided interval, a two-sided interval sym-

metric around the parameter, the shortest confidence interval, or some other interval might be preferable.

Comparing Two Normal Populations

In order to demonstrate the approach, we will show the case of comparing two normal populations. In the analysis of two-sample data, it is common to choose the t-test statistic to evaluate equality of the distributions. The test statistic is derived under the assumption of equal variances and independent normally distributed observations. We start by deriving the test statistic under this assumption, and later we derive a test variable when equality of variances is no longer assumed.

Let $X_1, \ldots, X_m$ be independent observations from a normal distribution $N(\mu_x, \sigma_x^2)$, and let $Y_1, \ldots, Y_n$, be independent observations from a normal distribution $N(\mu_y, \sigma_y^2)$. Then $\bar{X}$, $\bar{Y}$, S_x^2, and S_y^2 are the maximum likelihood estimators of μ_x, μ_y, σ_x^2, and σ_y^2, respectively. Since $\bar{X}$, $\bar{Y}$, S_x^2, and S_y^2 are complete sufficient statistics for the parameters of the two distributions, all inferences about the parameters can be based on them. The four statistics are independent, and their distributions are given by

$$\bar{X} \sim N(\mu_x, \frac{\sigma_x^2}{m}), \quad \bar{Y} \sim N(\mu_y, \frac{\sigma_y^2}{n}),$$

$$\frac{mS_x^2}{\sigma_x^2} \sim \chi_{m-1}^2, \quad \frac{nS_y^2}{\sigma_y^2} \sim \chi_{n-1}^2.$$

Under the assumption of equal variances ($\sigma^2 = \sigma_x^2 = \sigma_y^2$), inferences about the parameters can now be made on the basis of the complete sufficient statistics, $\bar{X}$, $\bar{Y}$, and

$$S^2 = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2}{m+n} = \frac{mS_x^2 + nS_y^2}{m+n}$$

and

$$\frac{(m+n)S^2}{\sigma^2} \sim \chi_{m+n-2}^2.$$

The parameter of primary interest is $\Delta = \mu_x - \mu_y$, and the hypotheses can be written as

$$H_0 : \Delta \leq 0 \text{ versus } H_a : \Delta > 0$$

or

$$H_0 : \mu_x \leq \mu_y \text{ versus } H_a : \mu_x > \mu_y.$$

The family of joint distributions of $\bar{X}$, $\bar{Y}$, and S^2 is both location- and scale-invariant, so we can reduce the problem to tests based on the statistic $T = (\bar{X} - \bar{Y})/S$. Because the distribution of $\bar{X} - \bar{Y}$ can be standardized as

$$\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(\phi, 1)$$

the distribution of T is given by

$$\frac{T\sqrt{mn(m+n-2)}}{\sqrt{m+n}} \sim t_{m+n-2\phi},$$

that is, the noncentral t-distribution with $m + n - 2$ degrees of freedom and the noncentrality parameter $\phi = \Delta/[\sigma\sqrt{1/m + 1/n}]$. The p-value is

$$P(T \geq (\bar{X} - \bar{Y})s^{-1} | \Delta = 0) = 1 - G_{m+n-2}((\bar{X} - \bar{Y})s^{-1}\sqrt{mn(m+n-2)/(m+n)}),$$

where s is the observed pooled standard deviation, and G_{m+n-2} is the cumulative distribution function of Student's t-distribution with $m + n - 2$ degrees of freedom.

It is well known that the t-test is the uniformly most powerful unbiased test for the situation above. The Wilcoxon rank-sum test is almost as efficient under these conditions (Lehmann 1975; Hodges and Lehmann 1956). If the distributions are heavy-tailed, the Wilcoxon rank-sum test is a more efficient test. When the alternative involves a change in scale as well as in location $F_x(t) = F_y((t - \Delta)/\sigma)$, then both these tests may be inefficient.

When the variances are not equal we are still interested in the inference about the difference $\Delta = \mu_x - \mu_y$. This problem has no exact fixed-level conventional test based on the complete sufficient statistics (Linnik 1968; Weerahandi 1994).

For instance, consider constructing interval estimates based on functions of the observed data. The difference in sample means is location-invariant, and its distribution is $\bar{X} - \bar{Y} \sim N(\Delta, \sigma_x^2/m + \sigma_y^2/n)$. The generalized pivotal quantity

$$R = (\bar{X} - \bar{Y} - \Delta)\sqrt{\frac{\sigma_x^2 s_x^2/(mS_x^2) + \sigma_y^2 s_y^2/(nS_y^2)}{\sigma_x^2/m + \sigma_y^2/n}}$$

can generate all invariant interval estimates '*similar*' in σ_x^2 and σ_y^2. Furthermore, let

$$Z = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\sigma_x^2/m + \sigma_y^2/n}}, \quad Y_x = mS_x^2/\sigma_x^2, \quad Y_y = nS_y^2/\sigma_y^2$$

where $Z \sim N(0, 1)$, $Y_x \sim \chi^2_{m-1}$, and $Y_y \sim \chi^2_{n-1}$ are all independent random variables. Moreover, the random variables $Y_x + Y_y \sim \chi^2_{m+n-2}$ and $B = Y_x/(Y_x + Y_y) \sim$

$Beta[(m-1)/2, (n-1)/2]$, and Z are also independently distributed. The pivotal quantity now becomes

$$R = Z\sqrt{s_x^2/Y_x + s_y^2/Y_y} = Z(Y_x + Y_y)\sqrt{s_x^2/B + s_y^2/(1-B)}.$$

Interval estimates of Δ based on R can be obtained from probability statements about R. The cumulative distribution function of R can be expressed as

$$P\{R \le r\} = P\left\{T \le r\sqrt{\frac{m+n-2}{s_x^2/B + s_y^2/(1-B)}}\right\} = EG_{m+n-2}\left\{r\sqrt{\frac{m+n-2}{s_x^2/B + s_y^2/(1-B)}}\right\}$$

where G_{m+n-2} is the cumulative distribution function of T and the expectation, E, is taken with respect to the beta random variable B.

The constant $c_\gamma = c_\gamma(s_x^2, s_y^2)$ needs to be found to satisfy

$$EG_{m+n-2}\left\{c_\gamma\sqrt{\frac{m+n-2}{s_x^2/B + s_y^2/(1-B)}}\right\} = \gamma.$$

A $100\gamma\%$ one-sided generalized confidence interval of Δ is $\left[(\bar{X}-\bar{Y}) - c_\gamma(s_x^2, s_y^2), \infty\right]$. A symmetric confidence interval about the point estimate $(\bar{X}-\bar{Y})$ of Δ is

$$(\bar{X}-\bar{Y}) - c_{(1+\gamma)/2}(s_x^2, s_y^2) \le \Delta \le (\bar{X}-\bar{Y}) + c_{(1+\gamma)/2}(s_x^2, s_y^2)$$

(Ogenstad 1998; Weerahandi 1994).

2.4 Illustrations

One-Way ANOVA Comparing Three Groups

Suppose that we have a dataset such that for comparing the mean effects of two active treatments (B and C) and a placebo (A). As can be experienced from analyzing a number of datasets, it is common that the variability in responses will increase with increasing mean levels. Let us say that after a preliminary review of the data and the figure we produced below (Fig. 2.1), based on equal sample sizes in the treatment groups, our '*intuition*' tells us that the treatment means are significantly different.

Although these data were indeed generated from normal populations with unequal means and variances, application of the classical F-test will not support our '*intuition*' in this case at all, because the p-value of the usual F-test is as large as 0.16. Using XPro (X-Technologies, Inc.), a software that calculates exact p-values, we compute the p-value for testing the equality of treatment means under the more reasonable assumption of unequal variances. The XPro Software produces a p-value that is 0.043, which is in line with the impression we get from the figure that we constructed. The discrepancy in p-values in this example is quite dramatic. It clearly demonstrates

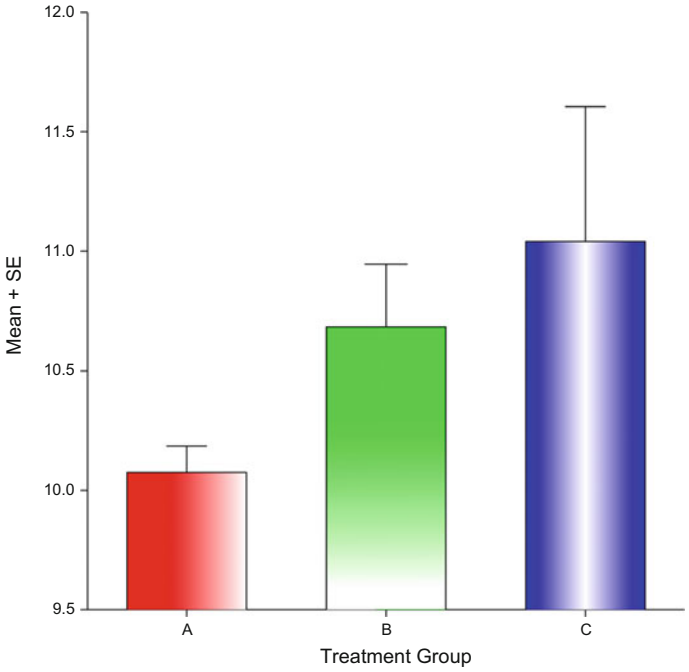


Fig. 2.1 Treatment group means + standard errors, based on equal sample sizes in the treatment groups

the serious weakness of the classical F-test in the presence of heteroscedasticity. Because the test ignores the problem of heteroscedasticity, the classical F-test fails to detect significant differences in treatments, despite the fact that the data provides sufficient information to do so. The complete ANOVA table to this illustration can be found in Appendix. As a note, the F-test is even more unreliable if the sample sizes in the treatment groups are different.

One-Way ANOVA Comparing Seven Groups

Although, based on equal sample sizes in the treatment groups, the treatment effects to the naked eye are quite different (Fig. 2.2), the p-value when applying the classical ANOVA to test the null hypothesis of equal means against the alternative hypothesis that not all means are equal is 0.11, which is not statistically significant at the 5% significance level. With the generalized F-test, the p-value without the equal variances assumption is 0.0098, which shows a very different outcome.

Repeated Measures Under Heteroscedasticity

We will now show an example of hemodynamic monitoring, which has long formed the cornerstone of heart failure (HF) and pulmonary hypertension diagnosis and management. There is a long history of invasive hemodynamic monitoring initially using pulmonary artery (PA) pressure catheters in the hospital setting, to evaluate the

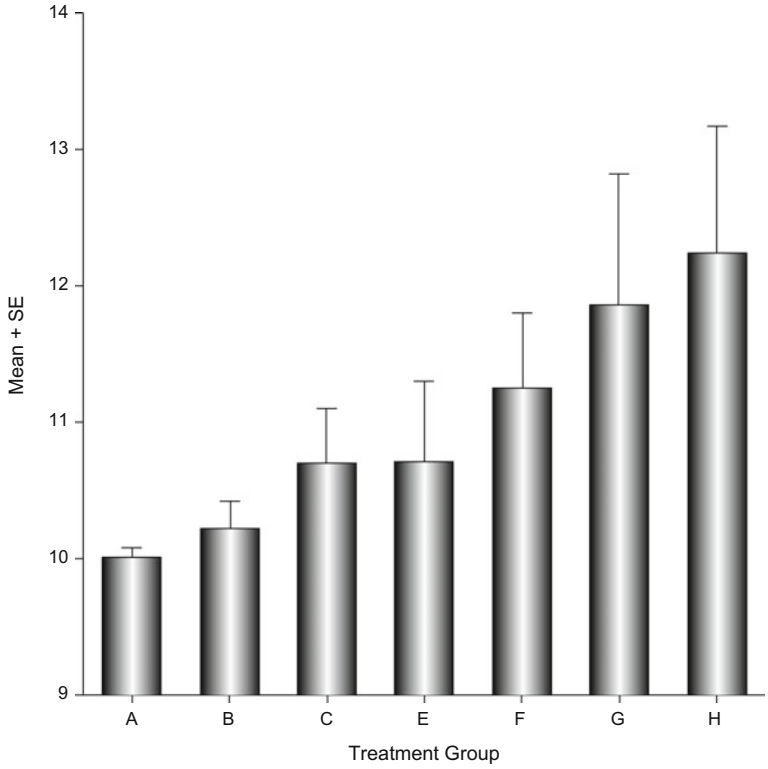


Fig. 2.2 Treatment group means + standard errors, based on equal sample sizes in the treatment groups

utility of a number of implantable devices that can allow for ambulatory determination of intracardiac pressures. Although the use of indwelling PA catheters has fallen out of favor in a number of settings, implantable devices have afforded clinicians an opportunity for objective determination of a patient's volume status and pulmonary pressures. Some devices, such as CardioMEMS' and thoracic impedance monitors present as part of implantable cardiac defibrillators, are supported by a body of evidence that show the potential to reduce HF-related morbidity and have received regulatory approval, whereas other devices have failed to show benefit and, in some cases, harm (Davey and Raina 2016).

We will consider potential data on pulmonary artery pressure where patients have been placed on one of four treatments ($G = 4$) to bring down the PA pressure. The patients have five scheduled visits at weeks 1, 2, 3, 4, and 5 with their investigator. Shown in Fig. 2.3 are bar graphs reflecting the arithmetic means, based on equal sample sizes in each group, with standard errors of a hypothetical dataset of normally distributed observations that was generated by simulating the following model

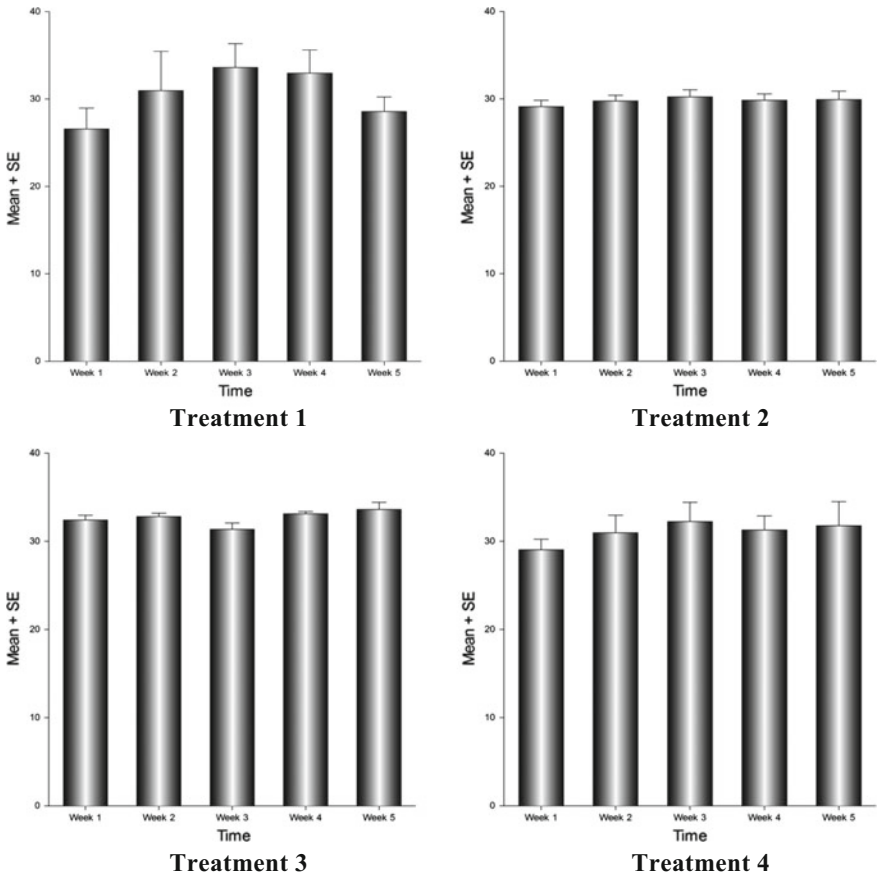


Fig. 2.3 Treatment group means + standard errors, based on equal sample sizes in the treatment groups and weeks

$$Y_{i(g)t} = \theta_g + \beta_t + \gamma_{gt} + \alpha_{i(g)} + \varepsilon_{i(g)t},$$

where $t = 1, \ldots, 5$, $i(g) = 1, \ldots, n_g$, $g = 1, \ldots, 4$. $\alpha_{i(g)}$ is the random effect due to among-subject variation, θ_g, $g = 1, \ldots, 4$ are the treatment effects, β_t, $t = 1, \ldots, 5$ are effects due to visits, γ_{gt} are their interactions, and ε_{it} are the residual terms.

Extending the usual assumption about variance components to possibly unequal group variances, we now have

$$\alpha_{i(g)} \sim N(0, \sigma_\alpha^2), \quad \varepsilon_{i(g)t} \sim N(0, \sigma_g^2),$$

where $t = 1, \ldots, 5$, $i(g) = 1, \ldots, n_g$, $g = 1, \ldots, 4$.

Although the data seems typical in a repeated measures design, a closer look at the data reveals that the treatment group variances, in this case, are substantially

Table 2.1 Classical analysis of variance results

ANOVA table					
Source	DF	SS	MS	F-value	P-value
Weeks	4	86.8247	21.7062	1.289	0.284
Treatments	3	110.992	36.9972	2.137	0.136
Within Treatment	16	276.997	17.3123		
Treatment × Weeks	12	135.048	11.254	0.668	0.775
Error	64	1078.08	16.845		
Total	99	1687.94			

different, which is evident in Fig. 2.3. Obviously, in this application, it is not reasonable to assume that the variances are equal. But should it make any difference to our conclusions whether or not the assumption is reasonable? To examine this, let us first ignore the fact that variances are different and apply the classical ANOVA as usually done by most people. The ANOVA table (Table 2.1) obtained by applying formulas for classical repeated measures analysis for the case of homoscedastic variances is shown below.

According to the p-values appearing in the ANOVA table, none of the effects including the treatment effect are significant. Now we will drop the equal variances assumption and retest the hypothesis that there is no difference in the mean PA pressures between the different treatments. The p-value for testing the difference between the treatments then becomes 0.0009. This means that the difference between the treatments is highly significant despite what the classical ANOVA suggested. Usually milder assumptions make the p-value of a test larger and power of a test smaller. But here the assumption of equal variances is so unreasonable that the p-value under the assumption of equal variances is substantially larger. This illustration clearly displays the reduction of the power of classical F-tests under heteroscedasticity.

2.5 Statistical Software

XPro computes exact p-values for testing hypotheses and computes confidence intervals based on exact probability statements. This becomes particularly important when one is using small or unbalanced data. The assumptions upon which standard methods are based are then typically biased, resulting in unrealistic p-values and confidence intervals. The software supports the exact inference in various linear models. It has been proven to be able to detect significant and nonsignificant experimental results early, even with small sample sizes. XPro procedures are complimentary to such program as StatXact which specialize in exact non-parametric methods, such as those dealing with contingency tables and categorical data. Most software programs

provide exact parametric methods only under the assumption of homoscedasticity in the ANOVA. In addition to such classical procedures, XPro provides procedures based on milder assumptions. To make this possible XPro performs high dimensional numerical integrations and solves highly nonlinear equations. The complexities of the underlying formulas make the problem of computing exact *p*-values and confidence limits very tedious. XPro makes use of efficient algorithms tailor made for exact inferences in linear models and provides an easy to use interface that facilitates all necessary analyses without passing the burden of any such numerical methods to the user. The methods used are based on Weerahandi (1994). *P*-values and confidence intervals, based on exact statistical calculations, are provided for a large number of following statistical procedures, models, and relationships.

As mentioned, StatXact (Cytel Corporation), is used for a host of nonparametric statistical procedures and sample size determination, and LogXact (Cytel Corporation), for the construction of logistic and Poisson regression models. Both StatXact and LogXact allow the user to select exact, Monte Carlo, or regular asymptotic methods of calculating *p*-values and confidence intervals. If exact methods take too long or are unavailable because of computer memory limitations, the user may select Monte Carlo techniques. Monte Carlo results are often very close to those produced by exact methods. XPro likewise provides the user with a Monte Carlo option for the majority of its procedures. It is generally used under the same conditions mentioned above.

Appendix

One-way ANOVA table

Sample sizes and MLEs of parameters

Column	Sample size	Sample mean	Sample variance
A	20	10.0752	0.237953
B	17	10.6838	1.08007
C	19	11.0419	5.66803

ANOVA Table

Source	DF	SS	MS	F-value
Treatment	2	9.3291	4.66455	1.88988
Error	53	130.813	2.46817	
Total	55	140.142		

Testing the Equality of All Means

Classical F-Test

P-value under the equal variances assumption: 0.161

Generalized F-Test

P-value without the equal variances assumption: 0.043

References

Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's *t* statistic under various non-normal distributions. *Journal of Educational and Statistics, 5,* 309–335.

Brownie, C., Boos, D. D., & Hughes-Oliver, J. (1990). Modifying the *t* and ANOVA *F*-tests when treatment is expected to increase variability relative to controls. *Biometrics, 46,* 259–266.

Burdick, R. K., Park, Y.-J., Montgomery, D. C., & Borror, C. M. (2005). Confidence intervals for misclassification rates in a gauge R&R study, *Journal of Quality Technology*.

Davey, R., & Raina, A. (2016). Hemodynamic monitoring in heart failure and pulmonary hypertension: From analog tracings to the digital age. *World Journal of Transplantation, 6*(3), 542–547.

Gamage, J., Mathew, T., & Weerahandi, S. (2013). Generalized prediction intervals for BLUPs in mixed models. *Journal of Multivariate Analysis*, *220*, 226–233.

Gamage, J., & Weerahandi, S. (1998). Size performance of some tests in one-way ANOVA. *Communications in Statistics Simulation and Computation, 27,* 625–640.

Ghosh, J. K. (1961). On the relation among shortest confidence intervals of different types. *Calcutta Statistical Association Bulletin* 147–152.

Good, P. (1994). *Permutation tests—a practical guide to resampling methods for testing hypotheses*. Springer.

Graubard, B. I., & Korn, E. L. (1987). Choice of column scores for testing independence in ordered $2 \times$ K contingency tables. *Biometrics, 43,* 471–476.

Hamada, M., & Weerahandi, S. (2000). Measurement system assessment via generalized inference. *Journal of Quality Technology, 32,* 241–253.

Hodges, I. L., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics, 27,* 324–335.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association, 72,* 789–808.

Kim, H. (2008). Moments of truncated Student-t distribution. *Journal of the Korean Statistical Society, 37,* 81–87.

Koschat, M. A., & Weerahandi, S. (1992). Chow-type tests under heteroscedasticity. *Journal of Business & Economic Statistics, 10*(22), 1–228.

Krishnamoorthy, K., Mathew, T., & Ramachandran, G. (2006). Generalized P-values and confidence intervals: A novel approach for analyzing lognormally distributed exposure data. *Journal of Occupational and Environmental Hygiene, 3,* 642–650.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks*. Oakland, CA: Holden-Day.

Linnik, Y. (1968). Statistical Problems with Nuisance Parameters. Translation of Mathematical mono-graph No. 20, American Mathematical Society, New York.

Meng, X. L. (1994). Posterior Predictive p-values. *The Annals of Statistics, 22*(3), 1142–1160.

Ogenstad, S. (1998). The Use of Generalized Tests in Medical Research. *Journal of Biopharmaceutical Statistics, 8*(4), 497–508.

Pratt, J. W. (1961). Length of confidence intervals. *Journal of the American Statistical Association, 56,* 541–567.

Thursby, J. G. (1992). A comparison of several exact and approximate tests for structural shift under heteroscedasticity. *Journal of Econometrics, 53,* 363–386.

Tsui, K., & Weerahandi, S. (1989). Generalized *p*-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association, 84,* 602–607.

Zhou, L., & Mathew, T. (1994). Some tests for variance components using generalized p-values. *Technometrics, 36,* 394–421.

Weerahandi, S. (1987). Testing regression equality with unequal variances. *Econometrica, 55,* 1211–1215.

Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association, 88,* 899–905.

Weerahandi, S. (1994). *Exact statistical methods for data analysis*. New York: Springer.
Weerahandi, S. (2004). *Generalized inference in repeated measure* (p. 44) New York: Wiley.
Weerahandi, S., & Tsui, K. (1996). Solving ANOVA problems by Bayesian approach, comment on posterior predictive assessment of model fitness via realized discrepancies by Gelman, Meng, and Stern, Statistica Sinica 6, 792–796.
X-Techniques, Inc (1994). XPro: Exact Procedures for Parametric Inference, X-Techniques, Inc, Millington, NJ.

Chapter 3
Discrete Time-to-Event and Rank-Based Methods with Application to Composite Endpoint for Assessing Evidence of Disease Activity

Macaulay Okwuokenye

3.1 Introduction

Many clinical trials include multiple objectives for evaluating efficacy and safety of therapies. These objectives are frequently addressed using multiple endpoints partly because most diseases have more than one consequence on biology. Multiple endpoints are used for evaluating therapies for several reasons: They enable assessment of consistency in treatment effects across different but critical features of a disease. They enable assessment of different features of same underlying pathophysiology of a disease. They characterize a disease with complex etiology better than a single endpoint (Hugue and Sankoh 1997) because a single endpoint rarely fully characterize clinical relevant benefits of a therapy. For example, in clinical trials for evaluating effects of disease modifying therapies in relapsing-remitting multiple sclerosis (RRMS), treatment benefits are evaluated in terms of reduction in relapse frequency, disability worsening, and number of magnetic resonance imaging (MRI) lesions. In clinical trials for evaluating effects of anti-neoplastic agents, treatment benefits are evaluated in terms of overall survival, progression-free survival, and shrinkage in tumor size by a specified amount. In clinical trials for evaluating therapies for epilepsy, treatment benefits can be evaluated in terms of time to first epileptic seizure and frequency of drop attacks.

Separate statistical analyses of multiple endpoints without appropriate adjustment for type I error attracts multiplicity of tests issues, and this could increase likelihood of incorrect conclusion. On the other hand, adjusting for type I error, using for example Bonferroni test, when the treatment effects on the component endpoints are small and the endpoints are many may be too conservative and will affect the statistical power of tests. Designing an adequately powered study to detect small treatment effects will not only require very large sample size, but also it will be costly and might require

M. Okwuokenye (✉)
Jiann-Ping Hsu College of Public Health, Georgia Southern University and University of New England, Syros Pharmaceutical, Cambridge, MA, USA
e-mail: mokwuokenye@syros.com; chiefendorce@hotmail.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_3

long time to recruit patients. When evaluating multiple endpoints on which treatment effects are similar and consistent, an approach that avoids multiplicity problem is the creation of a composite of these endpoints.

A composite endpoint is derived by collapsing multiple endpoints into a single value at the subject level, and using the single value as the unit for statistical analyses. In other worlds, it is the first occurrence of any of the component endpoints. A composite endpoint may be chosen in a trial to capture the presentation of a disease better than any single endpoint (Mascha and Sessler 2011); it is used to better characterize a disease that manifests in complex ways when no agreement exist among experts on the most relevant efficacy endpoint. It is used to possibly increase power when expected treatment effect is small and consistent across the individual endpoints through increase in number of events. Sample size for a clinical trial with a composite endpoint can considerably reduce for component endpoints that meaningfully contribute similar effects to overall treatment effects. A composite endpoint eases interpretation (Fairclouch 2010) when individual endpoints are of equal or similar importance. A composite endpoint helps avoid the problem of competing risks (Neaton et al. 2005).

This chapter presents statistical methods for analyzing a binary composite endpoint that is a function of continuous right censored time-to-event endpoints and another endpoint(s) whose exact time of occurrence is(are) unknown but only known to have occurred within an interval. In Sects. 3.2 and 3.3, respectively, the data structure and analyses methods, including estimation and inference based on discrete time-to-event, are described. Section 3.4 presents clinical trial example of the discrete time-to-event method; Sect. 3.5, some drawbacks of collapsed binary composite endpoint method of analyses. Section 3.6 presents a rank-based method for evaluating binary multiple endpoints.

3.2 Data Structure

Consider a clinical trial designed to compare efficacy of two therapies over a period of time. Suppose that for each treatment group, the study design calls for $K + 1$ clinic visits, and that the assessments were scheduled to occur at time-points $t = 0, 1, 2, \ldots, K$, representing weeks or months as dictated by the study protocol. For the endpoint with periodic assessment, let T be the time of assessment that an event was first observed to be present. For a subject that does not have a event through T, define $T = \zeta$, where $\zeta (> K)$ is some fixed integer. Hence, $T \in \{0, 1, 2 \ldots K, \zeta\}$. Denote by C the last time prior to when a subject drops out of the study or is lost to follow-up, and the subject was not known to have had the event; for subjects that do not dropout, $T = C$. Accordingly, the observed data for each subject is $\min(T, C)$.

3.3 Analyses Methods

3.3.1 Collapsed Binary Composite Endpoint

As noted in Sect. 3.1, the composite endpoint of interest is a function of continuous right-censored time-to-event endpoints and another endpoint(s) assessed periodically; hence, the exact time of occurrence for the latter is unknown but only known to have occurred within an interval.

3.3.1.1 Crude Incidence Rate

The crude incidence rate $P_e = n_e/N$ is the proportion that represents the number of patients with the composite endpoint (n_e) divided by the number of patients initially randomized (N) to the drug regardless of treatment duration (Kappos et al. 2011; Nixon et al. 2014). Although crude incidence rate is simple, it is only appropriate when used for short-term exposure to treatment, or when treatment duration is same for all patients, or when the composite endpoint occur close in time following treatment initiation.

3.3.1.2 Discrete Time-to-Event Method

As described in Okwuokenye (2015), a statistical analysis method proceeds by first organizing (grouping) the event times since after randomization for the continuous right-censored time-to-event endpoints into intervals, $A_i = [a_{i-1}, a_i)$, for $i = 1, \ldots, m$, where $a_0 = 0$ and $a_m = \infty$ with the event times in A_i recorded as t_i; the intervals are determined by the scheduled assessment visits of the periodically assessed endpoint (Okwuokenye 2015). Following grouping of the event times into intervals, a collapsed binary composite endpoint of any of the events versus none of the events is created to compare treatment regimen over the study period. A subject has the composite event if any of the component events occurs, and the composite event time is the minimum time of occurrence of any of the component endpoint. A discrete time-to-event (TTE) method is then applied to statistically analyze the composite endpoint.

Unlike the crude incidence rate analysis approach that ignores subjects' differential follow-up times and make unverifiable assumptions about event status of censored subjects, TTE approaches allow incorporation of subjects' differential follow-up times and appropriate handling of censoring. TTE methods allow appropriate weighting of loss to follow-up; therefore, subjects information are incorporated into the analyses for as long as they are known to be in the study. Additionally, beside allowing the patterns of event occurrence to be reflected, they ensure that statistical analyses are performed in the intention-to-treat population. They also enable

adjustment for time-dependent confounding and time-varying covariate for studies conducted over a long period.

3.3.1.3 Estimation Method for Composite Endpoint Proportion

After evaluating comparability of baseline characteristics, a suitable discrete time-to-event method may be applied to estimate and compare the proportion of subjects with the composite endpoint in the treatment groups using a reliability estimator, e.g., the Kaplan and Meier (1958) estimator or the actuarial estimators, e.g. life-table, of the probability of not having the composite endpoint over the study period. A difference between the actuarial and reliability method is how withdrawals are incorporated into the estimation of the conditional probability of the composite endpoint (events). The life-table and Kaplan-Meier (KM) methods make different assumptions about withdrawals. For tied events and loss to follow-up times, KM method assumes that all subjects lost to follow-up were at risk at the time of event. Number of withdrawals occurring in an interval are subtracted from the number at risk at the beginning of the interval. Life-table assumes that withdrawals occurred uniformly in an interval (Breslow and Crowley 1974). For tied events and lost to follow-up times, life-table assumes that half (1/2) of the subjects that were lost to follow-up were at risk at the time of events. Hence, in obtaining actuarial estimates, the number of withdrawals are halved before subtraction from number at risk at the beginning of the interval, as a protection against underestimation and overestimation of the proportion of subjects with the composite endpoint (Breslow and Crowley 1974).

Neither of the two approaches is superior to the other; the choice of estimation approach is driven by the assumption on dropouts and length of the intervals. For considerably wide intervals and unknown actual event times, the actuarial rather than reliability estimation may be more appropriate. Although the Kaplan-Meier estimate is the non-parametric maximum likelihood estimate of survival function relative to the class of all distributions (Kaplan and Meier 1958), for wide interval with grouped event times, it may be more reasonable to consider the number at risk for an interval to be the number at risk at the beginning of the interval minus half (1/2) the number of withdrawals during the interval. KM uses as the number at risk for an interval, the number at risk at the beginning of the interval minus the number of dropouts in that interval. KM method was developed for survival times on continuous scale with rare ties. The life-table method was developed for grouped data, where ties are more likely. The life-table method recognizes that it is unreasonable to assume that none of the dropouts was at risk in the interval and all were not at risk for the entire interval.

3.3.1.4 The Composite Endpoint Proportion

Let N_i denote the number at risk at the beginning of each ith interval, D_i the number known to have the event[1] in the interval, and W_i the number who discontinued in the interval not known to have had the event at time of discontinuation. The conditional probability of having the event in the interval via actuarial estimation method (standard life-table) is:

$$Q_i = \frac{D_i}{N_i - 1/2\,(W_i)}. \tag{3.1}$$

No concern here that Eq. (3.1) is undefined when $N_i = 0$ because practical settings typically have large sample. The conditional probability of not having the event in the interval is $P_i = 1 - Q_i$ given by:

$$\frac{N_i - D_i - 1/2\,(W_i)}{N_i - 1/2\,(W_i)} \tag{3.2}$$

The unconditional probability of not having the event is the product of the $P_i's$. For Kaplan-Meier method of estimation, $1/2\,(W_i)$ is dropped from Eq. (3.2). Not considering the number at risk to account for the withdrawals makes the $Q_i's$ smaller (due to larger denominator), leading to larger $P_i's$ and larger cumulative non-event rates. As the length of sub-intervals becomes smaller, the actuarial estimate of event probability approaches the Kaplan-Meier estimate as a limit, a reason the Kaplan-Meier is referred to as the product limit estimate (Cantor 2003, p. 21).

3.3.1.5 Comparing the Composite Endpoint Proportion

A hypothesis of interest is whether the composite endpoint proportions (or patterns) are equal (H_0) versus not equal (H_a). Write $H_0 : F_A = F_B$, versus any appropriate contradiction of the null, where F_A and F_B are the cumulative times to composite endpoint distributions for treatment groups A and B, respectively. Equivalently, H_0 may be stated in terms of $S = 1 - F$, where S represents the survival curve distributions. The reason for estimating the proportion having no composite endpoint is that the cumulative distribution function, which is the proportion with the composite endpoint, cannot be directly estimated in presence of censoring; therefore, it is estimated as $1 - S$, where S is the survival function.

Inference on the difference in proportions of subjects without the composite endpoint between two treatment groups may be obtained using non-parametric or parametric statistical methods. An example of a non-parametric method is the Cochran-Mantel-Haenzel (CMH) or Mantel-Cox approach (Mantel 1963; Mantel and Haenszel 1959; Mantel 1966).

[1] Event and composite endpoint are used interchangeable for ease of exposition.

Table 3.1 Data set-up for an interval for computing Mantel-Haenzel statistics

Treatment	Event	No event	At risk
A	n_{11i}	n_{12i}	n_{1+i}
B	n_{21i}	n_{22i}	n_{2+i}
	n_{+1i}	n_{+2i}	N_{++i}

Table 3.1 presents an example data structure for the ith interval. For an ith interval, let n_{11i} represent the number of subjects who have the composite endpoint and n_{12i} represent those without the composite endpoint among the number of subjects at risk n_{1+i} at the beginning of the interval in treatment group A. Similarly, let n_{21i} represent the number of subjects who have the composite endpoint and n_{22i} represent those without the composite endpoint among the number of subjects at risk n_{2+i} at the beginning of the interval in treatment group B. Let n_{+1i} represent number of subjects with the composite endpoint in the interval and n_{+2i} the number without the composite endpoint in the interval. Let $n_{1+i} + n_{2+i} + n_{+1i} + n_{+2i} = N_{++i}$.

The Cochran-Mantel-Haenzel statistics (CMH) (Mantel 1963) for assessing whether the time-to-composite endpoint pattern differs between the treatment groups A and B is:

$$\chi^2_{CMH} = \frac{\left[\left|\sum_i n_{11} - (n_{1+i} \times n_{+1i})/N_{++i}\right| - 0.5\right]^2}{V_{CMH}} \tag{3.3}$$

where $V_{CMH} = \sum_i (n_{1+i} \times n_{+1i} \times n_{2+i} \times n_{+2i})/(N^3_{++i} - N^2_{++i})$. Subtraction of 0.5 is a correction for continuity. See for example Mantel (1963) and Peace (2009) for a good exposition on utility of CMH for discrete time-to-event data.

Non-integer cell margins may arise when computing the effective sample size for standard life-table method. The CMH test is based on the assumption that the margins of the two-by-two tables is fixed and hence the distribution of the pivotal cell frequency is hypergeometric. Since the hypergeometric distribution applies to frequencies, (number of successes in n draws from a population of survival or cumulative probability of event without replacement) and not improper fractions, some will argue that CMH should not be applied to compare survival curves constructed by the life-table method. Nonetheless, if such an argument is of concern, one could use the discrete-time Cox's proportional hazards model to compare event rates in the two groups.

3.3.1.6 Assessing Covariate Effects

Cox's proportional hazard (Cox's PH) model (Cox 1972) is commonly used in assessing covariate effects when event times are continuous. The Cox's PH, specified in terms of hazard function $\lambda(t; X)$, is:

$$\lambda(t; \boldsymbol{X}) = \lambda_0(t) \exp(\boldsymbol{X}\boldsymbol{\beta}), \tag{3.4}$$

where $\boldsymbol{\beta}$ is a vector of p unknown parameters and $\lambda_0(t)$ is arbitrary. The hazard function for discrete failure time model (Hosmer and Lemeshow 1999; Prentice and Gloeckler 1978) can be derived from grouping event times in model (3.4) into disjoint intervals $A_i = [a_{i-1}, a_i)$, for $i = 1, \ldots, r$, where $a_0 = 0$ and $a_r = \infty$ with the event times in A_i recorded as t_i. For a subject with covariate vector $\boldsymbol{X}$, the probability of observing the composite event at time t_i is:

$$Pr(t_i|\boldsymbol{X}) = \{1 - \alpha_i^{\exp(\boldsymbol{X}(t_i)\boldsymbol{\beta})}\} \prod_{j}^{i-1} \alpha_i^{\exp(\boldsymbol{X}(t_i)\boldsymbol{\beta})}, \tag{3.5}$$

where

$$\alpha_i = \exp\{-\int_{a_{j-1}}^{a_j} \lambda_0(u)du\} \tag{3.6}$$

is the probability of composite event-free (i.e., not having the composite event) of a subject with $\boldsymbol{X}(t) = 0$ in interval A_j. The probability of surviving (i.e., been composite event-free) to the beginning of A_i is:

$$p(t_i; \boldsymbol{X}) = \prod_{j=1}^{i-1} \alpha_j^{\exp(\boldsymbol{X}(t_i)\boldsymbol{\beta})}. \tag{3.7}$$

The range of α_j is such that $0 < \alpha_j < 1$, $j = 1, 2, \ldots, r-1$; therefore, to remove restrictions on parameter range, α_j is substituted with the transformation $\phi_j = \log[1 - \log \alpha_j]$. So that the logarithm of the likelihood is:

$$ll = \delta \log[1 - \exp(1 - \exp(\phi_k + \boldsymbol{X}\boldsymbol{\beta}))] - \sum_{j-1}^{k-1} \exp(\phi_j + \boldsymbol{X}\boldsymbol{\beta}), \tag{3.8}$$

where $\delta = 1$ for observed event times; 0 for censored event times. With appropriate data structure, the discrete analog (Eq. 3.5) of the Cox's PH can be implemented in SAS using a binary regression model with the complementary log-log linearization transformation (Hosmer and Lemeshow 1999; Prentice and Gloeckler 1978). Some authors (e.g., Allison 1982) expressed the discrete-time model (3.5) describing association between T_i, the discrete failure time on an ith subject with covariate vector $\boldsymbol{X}_i$, as:

$$\ln[-\ln(1 - P_{ti})] = \boldsymbol{\beta}_{0t} + \boldsymbol{X}'_{ij}\boldsymbol{\beta}, \tag{3.9}$$

where $\boldsymbol{X}_{ij}$ is a $p \times 1$ covariate vector that may be constant over time or time-dependent (but fixed within a specific time interval), and $\boldsymbol{\beta}$ is a corresponding p-vector of parameters to be estimated. The function P_{ti} is the discrete hazard rate function given by

$$P_{it} = 1 - \exp(-\exp(\boldsymbol{\beta}_{0t} + \boldsymbol{X}'_{ij}\boldsymbol{\beta})).$$

The grouped discrete-time model in Eq. (3.5) assumes that the events are generated by Cox's PH (Prentice and Gloeckler 1978) and therefore yields hazard ratio estimates that are identical to those from the continuous time Cox's proportional hazard model. No assumptions are made about width of grouping interval in deriving marginal likelihood for $\boldsymbol{\beta}$ (Prentice and Gloeckler 1978; Allison 1982; Kalbfleisch and Prentice 1973). Although the discrete-time logistic regression model (Cox 1972) can be used for analyzing discrete failure times data, the parameter from such a model does not have relative risked interpretation, and the choice of grouping intervals impact the meaning of the regression coefficients (Prentice and Gloeckler 1978; Kalbfleisch and Prentice 1973).

3.3.2 *Summary Estimate Across Studies*

Pivotal proof of efficacy of a therapeutic agent typically requires two randomized placebo-controlled trials. To estimate the proportion of subjects having the composite endpoint from multiple studies, weighting may be applied to obtain pooled estimate of the proportions across studies. Denote by P_{ji} the probability of the composite endpoint for ith interval for study j and $P_{j\prime i}$ the probability of the composite endpoint for the ith interval for study $j\prime$. The pooled estimate for interval ith is:

$$\frac{N_{ji} \times P_{ji} + N_{j\prime i} \times P_{j\prime i}}{N_{ji} + N_{j\prime i}}, \quad j \neq j\prime \tag{3.10}$$

Equation (3.10) above implies weighting the individual study proportion of composite endpoint estimates for each interval proportionate to the number at risk for the interval from each study. The variance of the pooled estimate in Eq. (3.10) is:

$$f_{ji}^2 \times Var(P_{ji}) + f_{j\prime i}^2 \times Var(P_{j\prime i}), \tag{3.11}$$

where

$$fji = \frac{N_{ji}}{N_{ji} + N_{j\prime i}} \quad \text{for} \quad j = 1, 2, \ldots, J; \quad i = 1, 2, \ldots, I, \tag{3.12}$$

and $Var(P_{ji})$ is determined from the Greenwood (1926) formula give by:

$$Var\left(\hat{P}_{ji}\right) \approx \left[\hat{P}_{ji}\right]^2 \sum_{i=1}^{I} \frac{D_{ji}}{\left(N_{ji} - D_{ji}\right) \times D_{ji}}. \tag{3.13}$$

Other weights, such as inverse variance weight, may be used. For pooled estimate of treatment effects from multiple studies adjusting for covariates, one would first investigate treatment by study interaction using individual patient data (IPD) or meta-analysis (MA). If little or no evidence exits to suggest that treatment by study interaction is statistically significant, a fixed effects IPD model that blocks on study or a fixed effect MA model may be utilized. However, if evidence exists to suggest between study heterogeneity, then inference may be based on a random effects model. Similar conclusions are expected from both IPD and MA (Chen and Peace 2013). When using Cox's model to assess covariate effects, one could obtain summary estimate across study by stratification with study as the strata.

Once a study has successfully demonstrated effectiveness on a composite endpoint, other attributes of the composite endpoint should be analyzed. An example of such attributes are the treatment effects on individual endpoints. Analyses describing treatment effects on individual endpoints should accompany the results on treatment effects on composite endpoint. The results on individual endpoints will enable reviewers assess the extent to which the treatment effects on a composite endpoint are driven by any of the component endpoints.

3.4 Example

Multiple sclerosis is a chronic neurological disorder characterized by clinical and radio neurological disease activities. A treatment goal in the relapsing-remitting multiple sclerosis is the attainment of no evidence of disease activity (NEDA) following treatment for $t-$years. NEDA, now increasingly used for comparing disease modifying therapies, is no evidence of a clinical relapse, sustained disability worsening, new or enlarging T2 lesions, or T1 gadolinium-enhancing (Gd) lesions on MRI scan after a $t-$year exposure to treatment. Unlike relapses that have known onset dates of occurrence, MRI lesions dates of occurrence are unknown. What is typically known is that lesions are present or absent at the time of assessment, and that MRI lesions emerged between scheduled assessment visits.

Data in Table 3.2 are from a RRMS trial conducted over 48 weeks to compare two treatments in terms of no evidence of disease activity. In the trial, 180 patients were randomized into active drug; 182 subjects into placebo. MRI endpoints (Gd and T2 lesion) were evaluated at week 12, week 24, and week 48. In addition to MRI measures, subjects' onset dates of relapses was collected as was information about evidence of disability worsening. The interest was to estimate the proportion of subjects with no evidence of relapse, disability worsening, Gd, or T2 lesions after 48 weeks of treatment.

The event times of relapse and disability worsening were organized into grouped intervals, notably, (0–12], (12–24], (24–48) weeks, driven by the time of MRI assessment. The life-table method was applied to estimate the proportion of subjects without

Table 3.2 Life-table estimates of proportion of subjects without evidence of disease activity

Interval weeks	No event & discontinued	Events n (%)	Average # at Risk	P(No event)	SE(P)
Active drug					
(0–12]	10	143	175	0.1829	0.0292
(12–24]	0	5	27	0.1490	0.0275
(24–48)	1	8	21.5	0.0936	0.0232
Placebo					
(0–12]	11	129	176.5	0.2691	0.0334
(12–24]	3	9	40.5	0.2093	0.0314
(24–48)	0	9	30	0.1465	0.0281

Note Data are contrived for illustrative purpose only and do not represent any data from actual clinical trial. P(no event) and SE(P) are the life-table estimates of proportion of patients without the composite event and the corresponding standard error, respectively

evidence of disease activity. Life-table estimates of the proportion of subjects with no evidence of disease activity at week 104 are 9.36% and 14.65%, respectively, for active drug and placebo.

3.5 Drawbacks of Collapsed Binary Composite Endpoint

The collapsed binary composite endpoint makes it challenging to weight the component endpoints based on clinical importance. It implicitly assumes equal importance for each component endpoints leading to inadvertent overweighting of the component endpoint that occurs more frequent than others (Mascha and Sessler 2011). Estimate of treatment effects and results of test are driven by the component endpoint with the largest frequencies, with potential clouding of the component endpoint that occurred with less frequency. This is an issue when the overweighted component endpoint is of little clinical value. Additionally, inconsistent treatment effects could be challenging to interpret because treatment effects may defer for each of the component endpoint.

The collapsed binary composite endpoint does not discriminate between treatment groups in terms of number of component endpoints experienced by each patients. Consider six hypothetical subjects, three taking placebo and three taking active treatment. Suppose that two of the three subjects in the placebo have four of the component endpoints comprising binary composite endpoint and two subjects in the active treatment group have only one of the four component endpoints. Under the collapsed binary composite endpoint when patients are uncensored (for simplicity), the proportion of patients without the collapse binary composite endpoint and measure of disease activity are same; however, in the example above, a subject with only one of the component endpoint has different level of disease activity compared with a subject with all the four component endpoints. Such a difference in the level of disease activity can be reflected with a rank-based method.

3.6 Rank-Based Method

3.6.1 Ranking Binary Endpoints to Reflect Severity

When each of the component endpoint has a binary outcome, an expression for n-component endpoints each with binary outcome (a or b) is:

$$(a+b)^n = \sum_0^n \binom{n}{y} a^{n-y} b^y, \tag{3.14}$$

where n represents the number of component endpoints. Equation (3.14) is the Binomial Theorem, and the binomial coefficient $\binom{n}{y}$ represents the number of ways, without replacement, to choose y component endpoint(s) from a set of n endpoints. When no ordering exists in the occurrence of the events, these coefficients enable creation of a metric for severity of disease activity depending on disease activity combinations of the component endpoints. This metric induces categories or subgroups of severity of disease activity which could be ranked according to perceived severity. Considering the four component endpoints ($n = 4$) in the above RRMS trial example, this metric induces 16 categories. Under this metric, there are four categories for Gd lesion (Gd), T2 lesion (T2), clinical relapse (relapse) and sustained disability worsening only–taken one at a time; six categories for Gd, T2, clinical relapse or sustained disability worsening taken two at a time, etc. Surely, NEDA commands a rank of 1; the category obtained by taking all four at a time would command a rank of 16. Then ranks 2, 3, 4 and 5 would apply to the category taking one at a time, the ranks 6, 7, 8, 9, 10 and 11 would apply to the category taking two at a time, then ranks 12, 13, 14, 15 would apply to the category taken three at a time.

The assignment of the ranks may be accomplished with clinical input from project clinician. If the project clinician can not decide, the average rank of each subgroup is assigned to each member of the group. The rank assigned to each category should reflect the perceived severity of disease activity (SODA) score or evidence of disease activity (EDA) score. With this ranking method, one would still be able to estimate crude proportion of subjects with no evidence of disease activity, which is the proportion of subjects with SODA score of 1. Hence, SODA provides much more information than NEDA.

3.6.2 Statistical Analysis of the Ranks

If the data were presented as a 2 (treatment)-by-16 cross-tabulation, then one could compute the mean score (as well as the variance) for each row and statistically compare the rows. This can be done with the row mean scores resulting from the PROC FREQ with CHM option. The resulting test statistics would be a Chi-square

with one degree of freedom. For sufficiently large samples, the same inference above could be achieved with analysis of variance (Mack and Skillings 1980) using general linear model (GLM). In SAS this can be achieved with PROC GLM with the model statement

$$R_i = D_i + \epsilon_i, \tag{3.15}$$

where R_i and $D_i (i = 0, 1)$ are the severity scores and treatment indicator, respectively. The inference in Eq. (3.15) is based on the F-distribution with d degree of freedom, where d is the degree of freedom associated with the mean square error. The F-distribution with one degree of freedom for the numerator approaches the Chi-square with one degree of freedom when the denominator degree of freedom approaches infinity. The inference would be toward one treatment group having greater (or less) average severity than the other. If there is concern that distribution of the ranks is not normal to rely on inference using GLM, inference based on a permutation/randomization test or other appropriate non-parametric method may be pursued. Missing data are imputed prior to assigning ranks. This approach for comparing treatment effects on multiple endpoints assumes that there is no ordering in the events—and that occurrence of all the four events is more severe than the occurrence of three, which is more severe than occurrence of two, etc. Modification of the ranking is warranted if the events can happen in a particular order.

3.7 Concluding Remarks

In this chapter, we discussed the discrete failure times and the rank-based methods for analyzing evidence of disease activity. If a composite endpoint is of interest during study design, the study should be designed to allow for frequent assessment visits. To the extent possible, the component endpoints should be such that effects of treatment are similar and consistent across the component endpoints. Due to limited information, this might be difficult to determine a priori for a therapy been evaluated for new indication.

In the illustration above, implicit assumption in utilizing collapsed binary composite endpoint for analyses of NEDA is that the component endpoints are of approximately equal importance and of similar occurrence and consistent across the component endpoints. However, disability worsening tend to generally occur with less frequency, and the long term goal in the management of MS is to delay or prevent disability worsening; therefore, it would appear that disability worsening is one of the most important endpoints.

The rank-based method discussed in this chapter assumes complete data on all patients; therefore, missing data need to be imputed using appropriate data imputation method. To reflect event times of the composite endpoint on inference about treatment effects, weighted estimate may be used. An example of such a weight is a stabilized weight using time-to-composite endpoint. If the grouping intervals are small, then the weight could be based on Kaplan-Meir estimate or on life-table estimates for

more coarse interval. It is important that the assigned ranks reflect perceived severity of disease activity. If there is concern that the distance between the ranks might not reflect the same magnitude of disease severity, one could specify the lowest rank and assign value for the next rank based on perceived severity. One also could collapse closely related endpoints. For example, in the example above, the GD and T2 could be collapsed into one MRI endpoint, in which case this induces 8 categories; the category having relapse, disability worsening, and MRI will then command a rank of 8.

References

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, *13*, 61–98.

Breslow, N., & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, *2*(3), 437–453.

Cantor, A. A. (2003). *SAS Survival Analysis Techniques for Medical Research* (2nd ed.). NC, USA: SAS Institute Inc.

Chen, D.-G. D., & Peace, K. E. (2013). *Applied metal-analysis with R*. Chapman and Hall/CRC.

Cox, D. R. (1972). Regression models and life-tables (with discusion). *Journal of the Royal Statistical Society, Serries B*, *34*, 187–220.

Fairclouch, D. L. (2010). *Design and Analysis of Quality of Life Studies in Clinical Trials* (2nd ed.). Chapman and Hall/CRC.

Greenwood, M. (1926). The natural duration of cancer. In *Reports on public health and medical subjects* (Vol. 33, pp. 1–26). London: Her Majesty's Stationary Office.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied survival analysis: Regression modeling of time to event data*. Wiley.

Hugue, M. F., & Sankoh, A. J. (1997). A reviewer's perspective on multiple endpoint issues in clinical trials. *Journal of Biopharmaceutical Statistics*, *7*, 545–564.

Kalbfleisch, J. D., & Prentice, R. L. (1973). Marginal likelihood based on cox's regression and life model. *Bimetrika*, *60*(2), 267–278.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*, 457–481.

Kappos, L., Radue, E., & O'Connor, P., et al., (2011). Fingolimod treatment increases the proportion of patients who are free from disease activity in multiple sclerosis: Results from a phase 3, placebo-control study (freedoms). *Neurology*, *76* (Suppl 4, A563).

Mack, G. A., & Skillings, J. H. (1980). A friedman-type rank test for main effects in a two-factor ANOVA. *Journal of the American Statistical Association*, *75*, 947–951.

Mantel, N. (1963). Chi-square test with one degree of freedom: Extension of the mantel-haenszel procedure. *Journal of the American Statistical Association*, *58*, 690–700.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Report*, *50*, 163–170.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748.

Mascha, E. J., & Sessler, D. I. (2011). Design and analysis of studies with binary-event composite endpoints: Guidelines for anesthesia research. *Anesthesia and Analgesia*, *112*(6), 1461–1470.

Neaton, J. D., Gray, G., Zuckerman, B. D., & Konstam, M. A. (2005). Key issues in end point selection for heart failure trials: Composite end points. *Journal of Cardiac Failure*, *11*(8), 567–575.

Nixon, R., Tomic, N. B. D., Sfikas, N., Cutter, G., & Giovannoni, G. (2014). No evidence of disease activity: Indirect comparisons of oral therapies for the treatment of relapsing-remitting multiple sclerosis. *Advance Therapy*.

Okwuokenye, M. (2015). Discrete time-to-event and score-based methods with application to composite endpoint for assessing evidence of disease activity-free. In *Proceedings of Twenty-second (XXII) Biopharmaceutical Applied Statistics Symposium*, Rockville, Maryland, United States.

Peace, K. E. (Ed.). (2009). *Design and analysis of clinical trials with time-to-event endpoints*. Boca Raton, FL: Chapman and Hall/CRC.

Prentice, R. L., & Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, *34*(34), 57–67.

Chapter 4
Imputing Missing Data Using a Surrogate Biomarker: Analyzing the Incidence of Endometrial Hyperplasia

P. Lim and H. Jiang

4.1 Background

A low incidence of endometrial hyperplasia is a *sine qua non* for a marketing application of a new hormone replacement therapy (HRT). Biopsy and histopathological examination of the tissue specimen is the gold standard in diagnosing endometrial pathology associated with HRT. In published research of HRT, the non-invasive technique of transvaginal ultrasonography (TU) has been used to detect anatomical abnormalities of the endometrium—in parallel with timed biopsies. Endometrial thickness (ET) can be assessed ultrasonographically and a clinically useful correlation between ET and the absence or presence of endometrial hyperplasia as well as malignancy and other anatomical abnormalities has been described extensively (Osmers et al. 1990; Varner et al. 1991; Lin et al. 1991; Malpani et al. 1990; Granberg et al. 1991; Langer et al. 1997).

In a small percentage of cases, despite correct technical execution, endometrial biopsy fails to collect sufficient tissue material for histopathological diagnosis. One explanation is that an atrophic endometrium is hard to sample. Nevertheless, when a biopsy is classified as "insufficient tissue", its interpretation as *normal, atrophic endometrium* may be subject to suspicion when ET is 5 mm or more. ET as determined by TU can give additional assurance in such cases (FDA HRT Working Group 1995; CPMP 1997). In the pivotal clinical trials with the estradiol (E_2) and norgestimate (NGM) cyclophasic HRT product, it was decided to impute a hyperplasia likelihood in the cases that were classified as presenting "insufficient tissue" on end of treatment biopsy and where a coincident TU outcome was available.

P. Lim (✉)
Janssen Research & Development, LLC, Titusville, NJ, USA
e-mail: PLIM@its.jnj.com

H. Jiang
Janssen Research & Development, LLC, Raritan, NJ, USA

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_4

ET responds to HRT by showing an increase of 1 or 2 mm and most of this change happens over the first weeks. Also, hyperplasia incidence shows a detectable increase within 12 weeks of unopposed E_2 therapy. It was felt, therefore, that all subjects with steroid hormone exposure of 3 months or more could be included in a validation exercise.

The approach taken to validate TU outcomes for the purpose of imputing a likelihood of hyperplasia in cases of insufficient tissue resembles the evaluation of a *diagnostic test* against the *gold standard.* The validation exercise was limited to the specific subject population enrolled in prospective clinical trials with HRT. Results of two clinical trials in healthy postmenopausal subjects, which included on-treatment biopsies and coincident ET assessments by TU, were available for analysis. One was a Phase 2 study of an experimental HRT combination of E_2/NGM with E_2 doses of 1 and 2 mg and NGM doses of 0, 30, 90, and 180 μg. The other was a Phase 3 study of the same experimental HRT with E_2 (mg)/NGM (μg) doses of 1/0, 1/30, 1/90, and 1/180. In addition, data of comparable prospective clinical investigations from the published literature were reviewed. The test characteristics of the ET determination are described in terms of sensitivity and specificity versus the endometrial biopsy diagnosis. Sensitivity is the likelihood of correctly identifying hyperplasia; specificity is the likelihood of correctly excluding the presence of hyperplasia at a discrete ET.

No prospective attempts were made to re-evaluate precision and accuracy of ET assessments as TU is widely in use as part of the routine Obstetric/Gynecologic practice. Applying the technique as set out in the clinical trial protocols and investigator instructions, the ET in healthy postmenopausal women can easily be determined in discriminative steps of 1 mm.

4.2 Objective of the Study

In response to the guidelines issued by regulatory authorities, ET is utilized as a second line investigation to assess the likelihood of hyperplasia, given a case of on-therapy biopsy where insufficient tissue for diagnostic classification is obtained.

4.3 Validation

Three validation steps to evaluate the ET measurement characteristics to prospectively assess hyperplasia risk are undertaken:

1. investigate the influence of potential confounding factors. Data from the Phase 2 study were used.
2. determine the sensitivity and specificity of discrete ET. Data from the Phase 3 study were used.

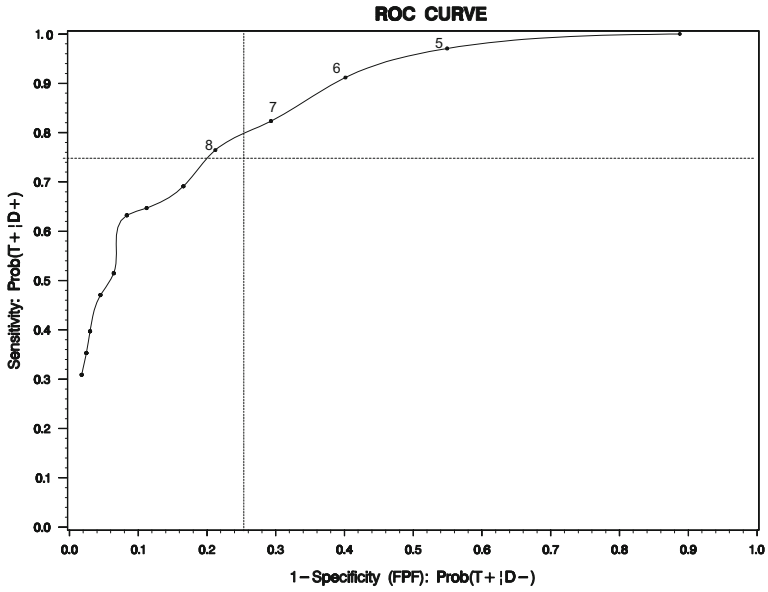


Fig. 4.1 ROC curve

3. evaluate the ET measurement characteristics derived from the E_2/NGM clinical trials against comparable data from published literature.

1. *Phase 2 Study*

To assess potential confounding factors, ET together with patient age, post-menopausal age, prior use of HRT and an indicator variable for continuous E_2 alone were investigated in a stepwise logistic regression analysis performed in the Phase 2 study. The final model consists of ET and the indicator variable for continuous E_2. Hence the only statistically significant predictor variables are the E_2 indicator and ET at end of treatment.

2. *Phase 3 study*

To determine the sensitivity and specificity of discrete 1 mm steps in ET, a receiver operator characteristic (ROC) curve was formed in all subjects in the Phase 3 study for whom biopsy and coincident ET results are available. The ROC curve of sensitivity (Y axis) versus (1-specificity) (X axis) is presented in Fig. 4.1. The optimum ET, exceeding 75% sensitivity while maximizing specificity, is 8 mm.

Using the threshold of 5 mm in order to increase sensitivity to detect hyperplasia, an imputed hyperplasia status was assigned for all subjects for whom biopsy and coincident ET results are available. Among the subjects with definitive hyperplasia readings based on biopsy, a logistic regression model where the probability of imputed hyperplasia is determined by treatment group showed that treatment was not significant. A similar logistic regression for subjects who had no hyperplasia

from definitive biopsy readings showed that treatment was a significant predictor. Hence, in adjusting the likelihood function for subjects with insufficient tissue readings, specificity is computed for each treatment group separately while sensitivity is computed across all treatment groups combined.

3. *Langer RD, Pierce, JJ, O'Hanlan KA et al.* (1997)

Most publications of HRT therapy are based on studies with a relatively small subject sample size and studies of an observational nature, i.e. case series or case control studies. Only the publication presenting the TU results obtained in the PEPI trial describes a prospective, randomized, double-blind design that is comparable with the E_2/NGM studies in terms of patient population, treatment and sample size.

The percentages of endometrial thickness in categories <5 mm, 5–10 mm and >10 mm are 42%, 45%, and 13%, respectively in the Phase 3 study. The ET distribution in no hyperplasia, hyperplasia and insufficient tissue cases are given below:

	No hyperplasia (N = 730)	Hyperplasia (N = 68)	Insufficient tissue (N = 47)
<5 mm	329 (45%)	2 (3%)	28 (60%)
5–10 mm	340 (47%)	23 (34%)	16 (34%)
>10 mm	61 (8%)	43 (63%)	3 (6%)

As expected the normal (no hyperplasia) and abnormal (hyperplasia) categories are not mutually exclusive. When 5 mm or more ET is used as cut-off for endometrial abnormalities, the similarities in test characteristics between E_2/NGM and PEPI trials are striking (numbers in parentheses are from the PEPI trial): sensitivity 97% (90%); specificity 45% (48%); positive predictive value 14% (9%); negative predictive value 99% (99%).

In conclusion, the validation steps described confirm ET at end of treatment as a predictor of hyperplasia risk with sensitivity and specificity numbers consistent with published literature. The dose of (unopposed) estrogen remains as the only independent determinant of hyperplasia incidence. As always, the choice of a cut-off value for ET reflects the compromise between sensitivity and specificity.

For the purpose of this exercise, it is felt that ET characteristics are satisfactory as a second line assessment to help estimate hyperplasia risk in cases of insufficient tissue. ET, corrected for misclassification by means of the likelihood function allows for imputation, because:

- the proportion of subjects with insufficient tissue is relatively small
- the ET in subjects with insufficient tissue is skewed to lower values, consistent with the expectation that atrophy rather than biopsy error is the cause
- the actual risk of hyperplasia is overestimated.

4.4 Analysis of Incidence of Endometrial Hyperplasia

To determine the incidence of hyperplasia in the Phase 3 study, each subject's on-treatment biopsy results are classified either as "no hyperplasia" or "hyperplasia." Any endometrial cancers have been included in the hyperplasia category. For each subject with insufficient tissue obtained ("no tissue obtained" or "no endometrial tissue obtained"), an imputed hyperplasia status is assigned based on the subject's endometrial thickness if obtained on the same day as the biopsy. A threshold value of 5 mm for ET is adopted based on its clinical appeal as presented above. Subjects with insufficient tissue from an on-treatment biopsy and an endometrial thickness greater or equal to the threshold (5 mm) are classified as having hyperplasia; those with endometrial thickness below the cut point are classified as having no hyperplasia.

A likelihood-based test is used to compare treatment groups. The log-likelihood function is the sum of the log-likelihood for all subjects with definitive biopsy data and the log-likelihood for subjects with insufficient tissue obtained. For subjects with definitive biopsy data, the likelihood is based on a logistic regression model where the probability of hyperplasia is determined by a single explanatory variable which takes on values 1, 2, 3, and 4, respectively, for treatments E_2/NGM 1/0, 1/30, 1/90, and 1/180. See Tukey et al. (1985). For subjects with insufficient tissue, the likelihood function further incorporates sensitivity and specificity to account for uncertainty in ultrasound measurement and other sources of error. A closed test procedure (Rom et al. 1994) for dose response was applied to maintain the overall significance level of 5%. Using this procedure, sequential testing was performed as follows: the homogeneity hypothesis of four successive proportions (Ho: $p_{1/0} = p_{1/30} = p_{1/90} = p_{1/180}$), the two homogeneity hypotheses of three successive proportions (Ho: $p_{1/0} = p_{1/30} = p_{1/90}$; Ho: $p_{1/30} = p_{1/90} = p_{1/180}$), and the three homogeneity hypotheses of two successive proportions (Ho: $p_{1/0} = p_{1/30}$; Ho: $p_{1/30} = p_{1/90}$; Ho: $p_{1/90} = p_{1/180}$ in this order). The reader is referred to the paper for the details of the procedure.

The primary analysis for endometrial histology included all subjects who had on treatment biopsies performed, whether or not the subjects completed 12 months of treatment. A secondary analysis included those subjects who had endometrial biopsies performed at month 12 of treatment or had endometrial hyperplasia diagnosed prior to month 12. Biopsies taken within one month after the completion of treatment were included in the analyses provided subjects have not received any other progestational or estrogenic treatment prior to the biopsy.

Two-sided 95% confidence intervals for the incidence of hyperplasia were provided for each treatment group using the likelihood ratio test.

See derivation of log likelihood ratio test at end of paper.

4.5 Results for the Phase 3 Study

Results of statistical analysis for the occurrence of endometrial hyperplasia are presented in Table 4.1 for all subjects with end-of-treatment biopsies. Subjects with insufficient tissue were not assumed to have inactive/atrophic tissue for the purpose of analysis; rather, an imputation was made of each subject's endometrial status (hyperplasia or no hyperplasia) based on the coincident ET measurement. If the ET was ≥5 mm, the subject was assigned to the category of hyperplasia, and if <5 mm, the subject was assigned to the category of no hyperplasia. In deriving the estimate and 95% confidence intervals in each group, the likelihood of hyperplasia for the imputed assignments was adjusted based on the sensitivity and specificity derived from definitive biopsy data.

In the population of subjects with end-of-treatment biopsies, the incidence of hyperplasia calculated for both the 1 mg E_2/90 μg NGM and 1 mg E_2/180 μg NGM groups was extremely low (5×10^{-4}% in each group), and the upper limit of the 95% confidence interval was less than 1%. In comparisons between treatment groups, highly significant differences ($p < 0.001$) were found for all comparisons shown in Table 4.1. It is noteworthy that the confidence intervals for the continuous 1 mg E_2 group, the 1 mg E_2/30 μg NGM and 1 mg E_2/90 μg NGM groups do not overlap, confirming the dose-related efficacy of norgestimate in preventing endometrial hyperplasia. The results clearly indicate that 90 μg is an adequate dose of norgestimate for administration with 1 mg E2 to protect the endometrium from the development of hyperplasia.

Acknowledgements The authors would like to acknowledge the contribution of Rosanne Lane (Janssen Research & Development), Qing Liu (USA), and Allan Sampson (University of Pittsburgh).

Derivation

Define {Y} to be the set of observations with definitive biopsy readings. The variable y is defined as 1, if the subject had hyperplasia; 0, otherwise. Hence, we can define p_i as follows:

$$p_i = P(y_i = 1) = e^{\alpha+\beta t_i}/(1 + e^{\alpha+\beta t_i}).$$

$$1 - p_i = 1/(1 + e^{\alpha+\beta t_i}).$$

This implies that logit $(p_i) = \alpha + \beta t_i$, where

$$t_i = \{1, \text{ if treatment is } E_2/\text{NGM } 1/0$$
$$= \{2, \text{ if treatment is } E_2/\text{NGM } 1/30$$

Table 4.1 Results of statistical analysis for the incidence of endometrial hyperplasia at the end of treatment (all subjects with biopsies performed after the start of treatment in the Phase 3 study)

Treatment group	Number of subjects in each biopsy category at the end of treatment				
	Hyperplasia	No hyperplasia	Insufficient tissue		
			Imputed hyperplasia	Imputed no hyperplasia	Not evaluable[a]
Continuous 1 mg E_2	63	176	4	4	1
Cyclophasic 1 mg E_2/30 μg NGM	13	222	8	7	1
Cyclophasic 1 mg E_2/90 μg NGM	0	219	2	9	4
Cyclophasic 1 mg E_2/180 μg NGM	0	220	5	8	2
	Estimate of the incidence of hyperplasia (95% C.I.)				
Continuous 1 mg E_2	26.114%		(20.848, 31.883%)		
Cyclophasic 1 mg E_2/30 μg NGM	5.467%		(3.046, 8.837%)		
Cyclophasic 1 mg E_2/90 μg NGM	0.0005%		(0, 0.848%)		
Cyclophasic 1 mg E_2/180 μg NGM	0.0005%		(0, 0.866%)		

Results of closed testing procedure using score test	
	p-value
Four regimens: 1/0, 1/30, 1/90, 1/180	<0.001
Three regimens: 1/0, 1/30, 1/90	<0.001
Three regimens: 1/30, 1/90, 1/180	<0.001
Two regimens: 1/0, 1/30	<0.001
Two regimens: 1/30, 1/90	<0.001
Two regimens: 1/90, 1/180	NS[b]

[a]Endometrial thickness measurement was not obtained on the same day as the biopsy
[b]Not significant

$$= \{3, \text{ if treatment is } E_2/\text{NGM } 1/90$$
$$= \{4\}, \text{ if treatment is } E_2/\text{NGM } 1/180.$$

Next, define {Z} to be the set of observations with insufficient tissue readings. The variable z is defined as 1, if hyperplasia was imputed (i.e., endometrial thickness ≥5 mm); 0, otherwise. Hence,

$$\begin{aligned} q_i &= \Pr(z_i = 1) \\ &= (c_i - d_i)\,[e^{\alpha+\beta t_i}/(1 + e^{\alpha+\beta t_i})] + d_i \end{aligned}$$

Note in the above formulation that $c_i = P(z_i = 1|y_i = 1)$ is sensitivity and $(1 - d_i) = P(z_i = 0|y_i = 0)$ is specificity.

The likelihood function is the product of the likelihood across all observations {y} (definitive biopsy readings) and {z} (insufficient tissue readings), as follows:

$$L(\alpha, \beta; Y, Z) = \prod p_i^{y_i}(1 - p_i)^{1-y_i} \quad \prod q_i^{z_i}(1 - q_i)^{1-z_i}$$

The log-likelihood is given as:

$$\begin{aligned} L &= \sum y_i \log p_i/(1 - p_i) + \sum \log(1 - p_i) + \sum z_i \log q_i + \sum (1 - z_i) \log (1 - q_i) \\ &= \sum y_i(\alpha + \beta t_i) + \sum \log(1 - p_i) + \sum z_i \log q_i + \sum (1 - z_i) \log (1 - q_i) \end{aligned}$$

The first derivatives of L with respect to α and β are:

$$\partial L/(\partial\alpha) = \sum y_i - \sum p_i + \sum \left[(z_i - q_i)/(q_i(1 - q_i)\right](c_i - d_i)\left[+e^{\alpha+\beta t_i}/(1 + e^{\alpha+\beta t_i})^2\right]$$

$$\partial L/(\partial\beta) = \sum y_i t_i - \sum t_i p_i + \sum \left[(z_i - q_i) \quad (c_i - d_i)/(q_i(1 - q_i)\right]\left[t_i e^{\alpha+\beta t_i}/((1 + e^{\alpha+\beta t_i})^2\right]$$

The second derivatives of L with respect to α and β are:

$$\begin{aligned} \partial^2 L/\partial\alpha^2 = &- \sum \partial p_i/\partial\alpha + \sum (ci - di)^2\left[\left(2Z_c q_i - Z_i - q_i^2\right)/\left(q_i^2(1 - q_i)^2\right]\right[\partial p_i/(\partial\alpha)]^2 \\ &+ \sum (ci - di)[(z_i - q_i)/(q_i(1 - q_i)\right](\partial^2 p_i/\partial\alpha^2) \end{aligned}$$

$$\begin{aligned} \partial^2 L/\partial\beta^2 = &- \sum t_i\,(\partial p_i/\partial\beta) + \sum (c_i - d_i)^2(\partial S_i/\partial q_i)\,(\partial p_i/\partial\beta)^2 \\ &+ \sum (ci - di)\,S_i\,(\partial^2 p_i/\partial\beta^2) \end{aligned}$$

$$\begin{aligned} \partial^2 L/\partial\alpha^2\partial\beta = &- \sum \partial P_i/\partial\beta + \sum (\partial S_i/\partial q_i)\,(c_i - d_i)^2(\partial p_i/\partial\beta)(\partial p_i/\partial\alpha) \\ &+ \sum (c_i - d_i)\,S_i\,(\partial^2 p_i/\partial\alpha\partial\beta) \end{aligned}$$

The terms comprising the above derivatives are defined below:

$$\partial p_i/\partial\alpha = \partial(1 - 1/1 + e^{\alpha+\beta t_i})/\partial\alpha = e^{\alpha+\beta t_i}/(1 + e^{\alpha+\beta t_i})^2$$

$$\partial p_i/\partial\beta = t_i e^{\alpha+\beta ti}/(1+e^{\alpha+\beta ti})^2$$

$$\partial^2 p_i/\partial\alpha^2 = [e^{\alpha+\beta ti} - e^{2(\alpha+\beta ti)}]/(1+e^{\alpha+\beta ti})^3$$

$$\partial^2 p_i/\partial\alpha\partial\beta = [t_i e^{(\alpha+\beta ti)} - t_i e^{2(\alpha+\beta\, ti)}]/(1+e^{\alpha+\beta ti})^3$$

$$\partial^2 p_i/\partial\beta^2 = [t_i^2 e^{\alpha+upbetati} - t_i^2 e^{2(\alpha+\beta ti)}]/(1+e^{\alpha+\beta ti})^3$$

$$\partial \log q_i/\partial\alpha = (c_i - d_i)\,(-e^{\alpha+\beta ti}/(1+e^{\alpha+\beta ti})^2)/\left[(c_i - d_i)p_i + d_i\right]$$

$$\partial \log q_i/\partial\beta = 1/\left(q_i\right)(c_i - d_i)\partial p_i/\partial\beta$$

$$\partial \log\left(1-q_i\right)/\partial\alpha = \left[1/\left(1-q_i\right)\right](-\partial q_i/\partial\alpha) = -\left[(c_i - d_i)/\left(1-q_i\right)\right]\partial p_i/\partial\alpha$$

$$\partial \log\left(1-q_i\right)/\partial\beta = \left[1/\left(1-q_i\right)\right](-\partial q_i/\partial\beta) = -\left[(c_i - d_i)/\left(1-q_i\right)\right]\partial p_i/\partial\beta$$

$$S_i = \left(z_i - q_i\right)/\left(q_i\left(1-q_i\right)\right)$$

$$q_i = (c_i - d_i)p_i + d_i$$

$$\partial S_i/\partial q_i = \left[2z_i q_i - z_i - q_i^2\right]/\left(q_i^2\left(1-q_i\right)^2\right)$$

$$\partial q_i/\partial\alpha = (c_i - d_i)\partial p_i/\partial\alpha$$

$$\partial q_i/\partial\beta = (c_i - d_i)\partial p_i/\partial\beta$$

$$\partial \log\left(1-p_i\right)/\partial\alpha = -e^{\alpha+\beta ti}/(1+e^{\alpha+\beta ti}) = -p_i$$

$$\partial \log\left(1-p_i\right)/\partial\beta = -t_i e^{\alpha+\beta ti}/(1+e^{\alpha+\beta ti}) = -t_i p_i$$

$$\partial \log\left(1 - p_i\right)/\partial \alpha^2 = -e^{\alpha+\beta t_i}/(1 + e^{\alpha+\beta t_i})^2$$

$$\partial \log\left(1 - p_i\right)/\partial \alpha \partial \beta = -t_i e^{\alpha+\beta t_i}/(1 + e^{\alpha+\beta t_i})^2$$

$$\partial \log\left(1 - p_i\right)/\partial \beta^2 = -t_i^2 e^{\alpha+\beta t_i}/(1 + e^{\alpha+\beta t_i})^2$$

No closed form expressions were obtained when the first derivatives were set equal to 0. Hence, the Newton-Raphson algorithm was used to compute the maximum likelihood estimates (MLEs) for α and β.

The score test was used for hypothesis testing as indicated in the closed testing procedure defined above. This is obtained by substituting $\beta = 0$ and forming the quantity

$$V_0' I_0^{-1} V_0$$

where V_0 is the vector of first derivatives and I_0 is the matrix of second derivatives.

To estimate the incidence for each treatment separately, the logistic model takes on a simple form and hence, it suffices to maximize the likelihood function with respect to α.

To construct the 95% confidence intervals, the following procedure was employed.

Let L_0 denote the value of the log likelihood substituting the MLE for P and L_1 for the log likelihood of any hypothesized P. Then $T = 2(L_0 - L_1)$ asymptotically follows the chi-square distribution χ_1^2 with one degree of freedom. The confidence interval is defined by the set of null hypotheses values for which we would not reject the hypothesis, that is, the set of values for which $1 - P(\chi_1^2 \leq T) - 0.05$ (or $P(\chi_1^2 > T) - 0.05$) becomes greater than 0.

References

Committee for Proprietary Medicinal Products (CPMP). (1997). Points to consider on hormone replacement therapy. EMEA Human Medicines Evaluation Unit, London, CPMP/EWP/021/97.

FDA HRT Working Group. (1995). Guidance for clinical evaluation of combination estrogen/progestogen-containing drug products used for hormone replacement therapy of post-menopausal women.

Granberg, S., Wikland, M., Karsson, B., et al. (1991). Endometrial thickness as measured by endovaginal ultrasonography for identifying endometrial abnormality. *American Journal of Obstetrics and Gynecology, 164,* 47–52.

Langer, R. D., Pierce, J. J., O'Hanlan, K. A., et al. (1997). Transvaginal ultrasonography compared with endometrial biopsy for the detection of endometrial disease. *The New England Journal of Medicine, 337*, 1792–1798.

Lin, M. C., Gosink, B. B., Wolf, S. I., et al. (1991). Endometrial thickness after menopause: Effect of hormone replacement. *Radiology*, *180*, 427–432.

Malpani, A., Singer, J., Wolverson, M. K., et al. (1990). Endometrial hyperplasia: Value of endometrial thickness in ultrosonographic diagnosis and clinical significance. *Journal of Clinical Ultrasound, 18*, 173–177.

Osmers, R., Volksen, M., & Schauer, A. (1990). Vaginosonography for early detection of endometrial carcinoma? *Lancet*, *335*, 1569–1571.

Rom, D. M., Costello, R. J., & Connell, L. T. (1994). On closed test procedures for dose response analysis. *Statistics in Medicine*, *13*, 1583–1596.

Tukey, J. W., Ciminera, J. L., & Heyse, J. F. (1985). Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics, 41,* 295–301.

Varner, R. E., Sparks, J. M., Cameron, C. D., et al. (1991). Transvaginal sonography of the endometrium in postmenopausal women. *Obstetrics & Gynecology*, *78*, 195–199.

Chapter 5
Advancing Interpretation of Patient-Reported Outcomes

Joseph C. Cappelleri and Andrew G. Bushmakin

5.1 Introduction

A patient-reported outcome (PRO) is any report on the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else (Food and Drug Administration 2009). As an umbrella term, PROs include a whole host of subjective concepts such as pain, fatigue, depression, aspects of well-being (e.g., physical, functional, psychological), treatment satisfaction, health-related quality of life, and physical symptoms such as nausea and vomiting. Patient-reported outcomes are often relevant in studying a variety of conditions—including pain, erectile dysfunction, fatigue, migraine, mental functioning, physical functioning, and depression—that cannot be assessed adequately without a patient's evaluation and whose key questions require patient's input on the impact of a disease or its treatment (McLeod et al. 2018).

To be useful to patients and other decision makers (e.g., physicians, regulatory agencies, reimbursement authorities), who are stakeholders in medical care, a PRO measure must undergo a validation process to confirm that it is reliably measuring what it is intended to measure. As assessments of subjective concepts, therefore, PRO measures require evidence of their validity (the instrument measures what it is intended to measure) and reliability (scores are stable and reproducible when they should be) before they can be used with confidence (Cappelleri et al. 2013; de Vet et al. 2011; Fayers and Machin 2016; McLeod et al. 2018; Streiner et al. 2015).

Given that a PRO measure has evidenced validity and reliability, this chapter addresses interpretation of PRO measures by advancing and enriching what their

J. C. Cappelleri (✉) · A. G. Bushmakin
Pfizer Inc, 445 Eastern Point Road, MS 8260-2502, Groton, CT 06340, USA
e-mail: joseph.c.cappelleri@pfizer.com

A. G. Bushmakin
e-mail: andrew.g.bushmakin@pfizer.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_5

scores mean for better understanding by stakeholders such as patients and their families, clinicians, researchers, payers, and regulators. Section 5.2 covers anchor-based approaches, Sect. 5.3 targets distribution-based approaches, and Sect. 5.4 highlights mediation models. Section 5.5 provides a summary. Throughout the chapter, concepts are illuminated with illustrative and real-life examples.

5.2 Anchor-Based Approaches

An anchor is a measure or criterion related to the targeted PRO under examination (Guyatt et al. 2003). As defined in this chapter, an anchor can be a measure different from or part of the PRO measure under consideration. The chosen anchor should be clearly understood in context and be easier to interpret than the PRO measure of interest, and the anchor should be appreciably or moderately correlated with the targeted PRO. An anchor-based approach links the targeted concept of the PRO measure to the meaningful concept or criterion emanating from the anchor.

Anchor-based approaches are the preferred way to enhance the clinical interpretation to the targeted PRO measure. They link the targeted PRO instrument under consideration with an anchor measure or indicator that is interpretable itself or lends itself to interpretation. Considerations for anchor-based methods include the nature of the relationship (e.g., linear) between the anchor and targeted PRO measures, the type of anchor, and the study population of interest. Several variants of anchor-based approaches are available (Crosby et al. 2003; Revicki et al. 2007; Fayers and Machin 2016). What follows are five types of anchor-based methods (Cappelleri et al. 2013).

5.2.1 Percentages Based on Thresholds

One of the simplest forms of presentation and interpretation is to show the percentage of patients above and below some specified value, which is an anchored value with a meaningful criterion (Fayers and Machin 2016). The method of percentages based on thresholds can be useful when the thresholds on a PRO measure are chosen judiciously so that their values have relevance, rather than being some arbitrary cut point. For example, a score above 25 on the erectile function domain of the International Index of Erectile Function is regarded as having normal erectile function (Cappelleri et al. 1999). In comparative studies, the proportion of patients in each treatment group who fall into this normal category can be noted and compared.

Establishing thresholds of a PRO on disease severity levels is another example of the use of an anchor-based approach. Such was the case when disease severity levels were obtained on the Fibromyalgia Impact Questionnaire (FIQ), a disease-specific composite developed to capture the spectrum of problems related to fibromyalgia and responses to therapeutic intervention (Bennett et al. 2009).

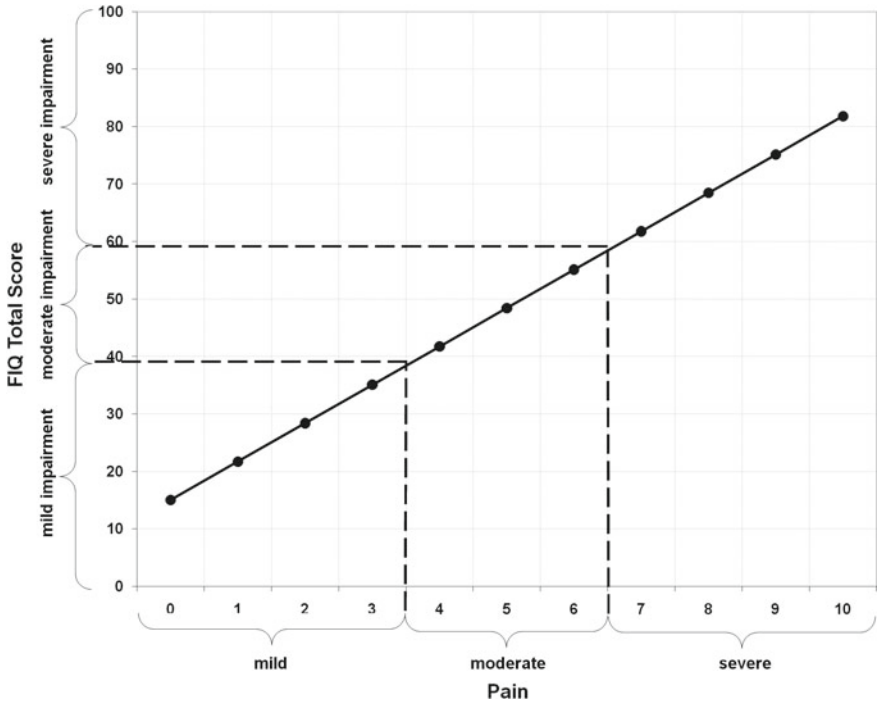


Fig. 5.1 Severity categorization of FIQ total score using pain severity as an anchor. Source: Cappelleri et al. (2013)

As strong Pearson correlations (i.e., an average of 0.67) were found between the FIQ total score and the average pain scores, it was reasonable to determine FIQ severity categories scores using pain severity as an anchor. A FIQ severity categorization was created corresponding to values of 3.5 and 6.5 on the pain scale taken as prespecified boundaries between pain severity categories (averaging pain over time transforms the original integer values from 0 to 10 to a continuous variable from 0 to 10). A repeated-measures model was used to estimate the relationship between the FIQ total score (outcome) and average pain scores (predictor) assessed at pre-treatment and post-treatment. A FIQ total score from 0 to<39 was found to represent a mild impact, ≥39 to <59 a moderate impact, and ≥59 to 100 a severe impact (Fig. 5.1). The severity bands can be useful in assessing treatment differences, as a criterion for study inclusion, and even to serve as an anchor to define clinically important differences for other patient-reported outcomes.

5.2.2 Criterion-Group Interpretation

A criterion-group interpretation, which is related to interpretation based on threshold percentages, involves the comparison of scores from the particular group of interest to a criterion group, a known group worthy of comparison which can serve as a yardstick for interpretation. This method of interpretation requires that meaningfully different groups be defined in the setting of interest. As a result, practitioners and other informed observers will readily comprehend the practical or clinical significance of the burden of illness or the treatment effect generated from contrasting the well-defined and distinct groups. Thus, the method requires in part a consensus of how meaningfully different groups are defined and who they are.

In a variation of criterion-group interpretation, PROs may use population-based reference values, which provide expected or typical scores that are called norms, to benchmark or anchor interpretation on PROs in the disease population of interest (Marquis et al. 2004; Fayers and Machin 2016). Baseline scores on the Medical Outcomes Study (MOS) Sleep Scale from two trials for the treatment of fibromyalgia were compared with scores obtained from a nationally representative sample in the United States (Cappelleri et al. 2009). Higher scores on the MOS Sleep Scale indicate more of the attribute being assessed (e.g., more sleep disturbance, more sleep adequacy). A one-sample z-test for the mean was also conducted to test whether the mean of each subscale from each of the two trials differed statistically from the corresponding normative mean, taken as a fixed targeted value.

Scores for each subscale of the MOS Sleep Scale were statistically ($P<0.001$) and substantially poorer than the general population normative values in the U.S., suggesting that patients with fibromyalgia have greater sleep problems relative to the general population (Fig. 5.2). For instance, patients with fibromyalgia reported sleeping an average of 5.4 and 5.6 h per night in the two studies, while the general population reported an average of 6.8 h of sleep per night.

5.2.3 Content-Based Interpretation

A content-based interpretation of a multi-item PRO scale uses a representative item, along with its response categories, internal to the measure itself to understand the meaning of different scores on that measure (Ware et al. 2007). In addition to descriptive statistics, item response theory (Chang and Reeve 2005), ordinal logistic regression (O'Connell 2005) and binary logistic regression (Kleinbaum and Klein 2010) can be used for content-based interpretation.

For example, a content-based interpretation was applied using the Rasch model, a type of item response theory model, on the six-item near-vision subscale of the German version of the 39-item National Eye Institute Visual Function Questionnaire in 200 patients with age-related macular degeneration (Thompson et al. 2007). Scores ranged from 0 (worst) to 100 (best). For a given subscale score, an estimated

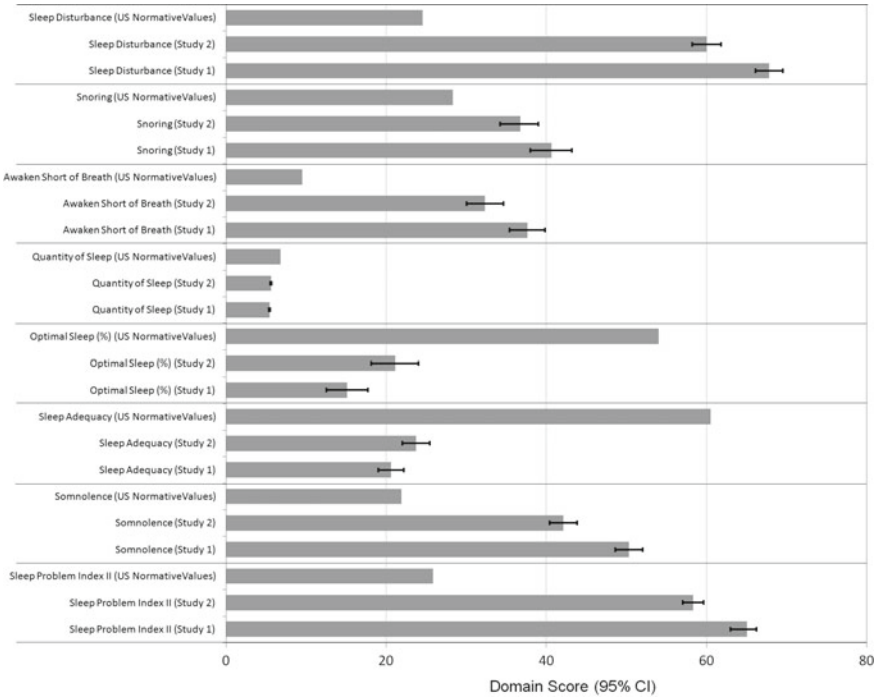


Fig. 5.2 Baseline mean scores (95% confidence intervals) on the medical outcomes study sleep scale for patients with fibromyalgia versus values from the United States general population. Source: Cappelleri et al. (2013)

probability of responding to each category of an ordinal item was obtained and the probabilities of responding to the two most favorable categories (no difficulty or little difficulty) were combined.

An individual, for instance, with an estimated true score of 75 on the near-vision subscale was expected to have approximately a 27% chance of little or no difficulty with reading small print, a 94% chance of little or no difficulty with finding an object on a crowded shelf, and nearly full certainty of little or no difficulty with shaving/styling hair/applying makeup (Fig. 5.3).

5.2.4 Clinically Important Difference

Highly significant *p*-values indicate little about the magnitude of a difference; statistical significance does not imply clinical significance. While a small p-value makes it likely that a real difference exists, which may be driven solely by an adequate or

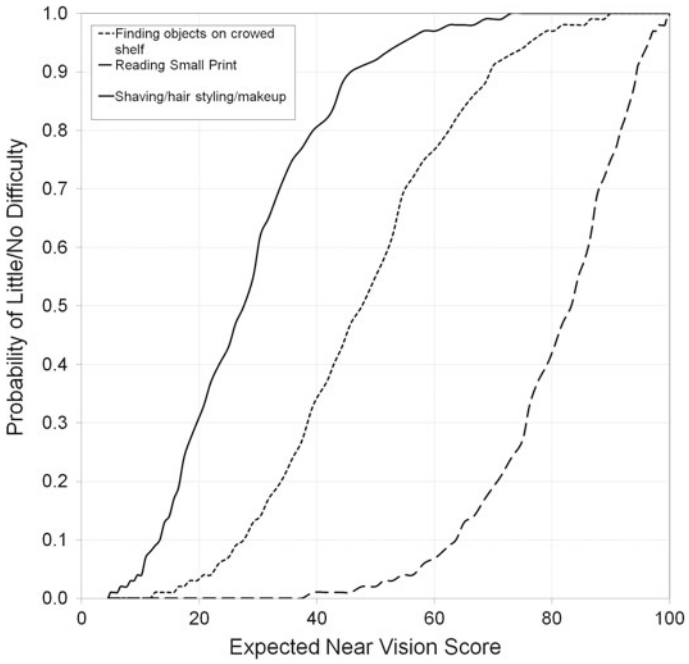


Fig. 5.3 Probability of little or no difficulty on three illustrative items from near-vision subscale of the National Eye Institute Visual Function Questionnaire Source: Cappelleri et al. (2013)

a large sample size, the observed or real difference might in fact not be clinically relevant.

A clinically important difference is the difference in scores between two treatment groups that can be considered clinically relevant (Coon and Cappelleri 2016). An anchor-based approach to quantify a clinically important difference (CID) on a PRO scale involves the use of an external measure—the anchor—that is clearly interpretable and is appreciably correlated with the targeted PRO measure (Guyatt et al. 2003). Responses on such an anchor can come from clinical measurements, clinician report, observer report, or, preferably, patient report. For example, patients can be asked to rate the extent of their change in their overall health status retrospectively since the beginning of the study on a 7-point scale [Patient Global Impression of Change (PGIC)]: 1 = "very much improved," 2 = "much improved," 3 = "minimally improved," 4 = "no change," 5 = "minimally worse," 6 = "much worse," 7 = "very much worse." Then the mean changes from baseline on the PRO scale of interest can be obtained for each of the categories on the anchor, and the differences in mean changes on the PRO scale between adjacent categories on the anchor can be examined for a clinically important difference. Data used in the analysis would be pooled across all treatments.

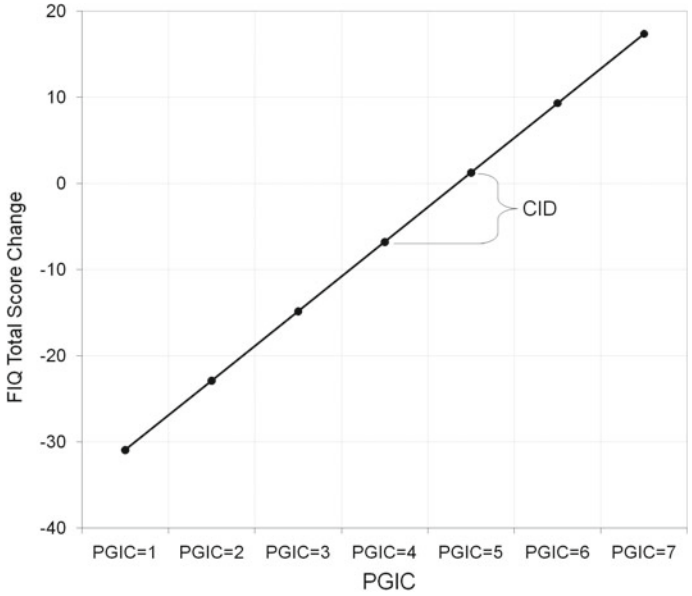


Fig. 5.4 Estimate of clinically important difference on the FIQ total score using PGIC as an anchor. *FIQ* Fibromyalgia Impact Questionnaire; *CID* clinically important difference; *PGIC* Patient Global Impression of Change. Source: Cappelleri et al. (2013)

As an example, reconsider the FIQ for patients with fibromyalgia (Bennett et al. 2009). A repeated measures model was used to estimate the mean change from baseline in FIQ total scores (range: 0–100 points) corresponding to each category on the PGIC, the anchor taken as a continuous predictor (Fig. 5.4). Differences in these mean changes between adjacent categories of PGIC corresponded to a clinically important difference of 8.1 [95% confidence interval (CI): 7.6–8.5].

An alternative or addition to PGIC, which is asked only at follow-up, a question can be asked serially—for example, one at baseline and one at follow-up—about the severity of a patient's overall current condition (e.g., none, mild, moderate, severe), and the difference in the mean PRO scores between adjacent categories on the anchor can be examined for a clinically important difference on the PRO scale (data should also be pooled across treatments). Using such a serial anchor, which focuses on the current state (either at the time when asked or over a relatively short time frame until the present), may address potential recall issues that may arise from a retrospective assessment, which require patients to compare their current state retrospectively relative to the start of the study.

As an example, a repeated-measures longitudinal model using all available data was used to estimate the relationship between Itch Severity Score (ISS), scored from 0 ("no itching") to 10 ("worst possible itching") on a 11-point numeric rating scale, and patient global assessment (PtGA) that was used as a continuous anchor predictor in a study on patients with psoriasis (Mamolo et al. 2015). The PtGA, which evaluates the

overall extent of cutaneous disease at a given time, had categories of "clear," "almost clear," "mild," "moderate," and "severe." It was assessed at baseline and weeks 4, 8, 12, and 16. A CID on the ISS was defined as the difference corresponding to a one-category change on the PtGA. The CID on ISS (95% CI) was estimated to be 1.64 (1.50–1.78). By Day 10, the mean change from baseline in ISS values for doses of the active treatment (placebo-adjusted) exceeded 1.64. A sensitivity analysis with PtGA as a categorical anchor predictor (instead of a continuous one) gave a similar estimate on CID and therefore supported the linearity assumption imposed by the model when PtGA was taken as a continuous anchor.

An estimated CID may vary in different situations because of natural sampling variation, different study populations, type of anchor or external criterion, time of assessment, and other considerations. For the same reasons, a CID is not necessarily a minimally clinically important difference, which is a more challenging avenue to pursue (Copay et al. 2007; King 2011; McLeod et al. 2016).

5.2.5 Clinically Important Responder

According to the FDA final guidance on PROs for a label claim, it is recommended to display individual responses using a priori responder definition: the threshold value on an individual observed PRO change score (be it the absolute or percent) that is to be interpreted as a treatment benefit (Food and Drug Administration 2009; McLeod et al. 2011). The proportion of subjects meeting the responder definition can then be reported for each treatment group and compared between groups. The responder definition can be determined empirically using an anchor-based approach.

While the CID refers to the difference in scores between two treatment groups that are considered clinically relevant, the clinically important responder (CIR) refers to the amount of change an individual patient would have to report to indicate that a relevant treatment benefit has been experienced (Coon and Cappelleri 2016). Hence the CID is considered a group-level interpretation, whereas CIR is considered an individual-level interpretation.

As an example, reconsider the ISS, this time in terms of its CIR threshold. In the same study reported previously on patients with psoriasis in Sect. 2.4 (Mamolo et al. 2015), the anchor Subject Global Impression of Change (SGIC) was created with three categories, consistent with FDA guidance on PRO measures (Food and Drug Administration 2009). Change from baseline on PtGA was used to generate the SGIC. If post-baseline PtGA improved relative to baseline PtGA, SGIC was defined as "better" (with a value of −1); if PtGA worsened relative to baseline PtGA, SGIC was defined as "worse" (1); and if PtGA was unchanged relative to baseline, SGIC was defined as "same" (0). A repeated-measures model was used to estimate the relationship between percent change from baseline ISS and the SGIC as an anchor. The difference in percent changes on the ISS corresponding to a one-category change on the SGIC was used to define the CIR threshold.

The ISS CIR (95% CI) was estimated to be 29.85% (23.30–36.40). A 30% improvement on the ISS was therefore used to define a responder. At Week 2, the proportion of responders was 77.8, 68.8, and 76.6% for the three doses of the active intervention versus 34.0% for placebo. This improvement was sustained through Week 12, when the percentage of responders increased to 91.9, 87.2, and 100.0% for three doses of the active intervention versus 29.4% for placebo; therefore, relative to placebo, the active intervention gave correspondingly an excess percentage of responders of 62.5, 57.8 and 70.6%.

A sensitivity analysis, with SGIC taken as the categorical predictor (instead of a continuous predictor), confirmed that the distance of "better" and "worse" was approximately symmetric around "same," thereby supporting SGIS as a continuous predictor. If such a sensitivity analysis instead showed asymmetry with "better" versus "same" not being equidistant with "same" versus "worse," the larger of the two differences could be used instead (to be conservative).

Responder analysis is a determined attempt to understand whether the effect of an intervention, shown to be statistically significant on a PRO measurement scale, has clinical significance. While it has defenders (Lewis 2004), its limitations have been reported (Snapinn and Jiang 2007). Limitations include the expected reduction in statistical power when moving from a continuous to binary outcome and, when not assessed empirically and justifiably, the potential for an arbitrary cutoff score to bifurcate or separate responders from non-responders. Responder analysis is best positioned as a descriptive display and as an adjunct to—as a complement and supplement to—the main analysis based on the full original scale of measurement using established statistical methods (e.g., repeated measures or random coefficient models when the data are longitudinal). As is the case for CID, a value for CIR is not necessarily a minimum threshold.

5.3 Distribution-Based Approaches

The methods highlighted thus far to estimate the magnitude and meaning of important changes or difference make use of clinically-based, patient-centered information on an anchor measure related to the PRO scale of interest. To complement such information, approaches based strictly on the distribution of the data may prove insightful. Such distribution-based methods can offer valuable insights about the magnitude of an effect. These methods also allow for a standardization of different scales with different ranges and ways of scoring. On the other hand, a limitation of distribution-based methods should be noted: although their interpretation can be meaningful, they do not provide information about *clinical* meaningfulness (Hays et al. 2005). Several types of distribution-based metrics are available (Crosby et al. 2003; Revicki et al. 2007; Fayers and Machin 2016). In this section a few of them are highlighted (Cappelleri et al. 2013).

5.3.1 *Effect Size*

The effect size (ES) can be an informative metric to gauge the magnitude of differences in a PRO scale within a group or between two groups. As presented here, we define the ES represents any standardized metric with difference in means in its numerator and a measure of variability in its denominator. The (standardized) effect size is a kind of signal-to-noise ratio that quantifies the heft or amount of effect relative to variability. Variations on this type of ES exist based on what is taken as a measure of variability in the denominator (Fayers and Machin 2016).

When standard deviations differ in the denominator of the ES, results from different studies for the same PRO metric can give different ES values even though their differences in means (numerators) may be the same. The ES should therefore be accompanied by its constituent elements, namely, its means, standard deviations and sample sizes.

Values of ES from different scales on the same intervention render standardized changes whose magnitudes can be fairly compared on the same dimensionless scale, despite the scales having different ranges of values. In addition, ES provides a general set of thresholds or benchmarks on the impact of an intervention, with values of 0.2 standard deviation units generally regarded as "small," 0.5 as "medium," and 0.8 as "large" (Cohen 1988).

For the impact of an intervention within a single group, ES values have commonly appeared in at least two ways. One way is the mean of changes in the scores recorded by the same subjects at two different times, divided by the standard deviation of these changes in scores. Note that the standard deviation in this case will be affected by the effects of the intervention over time, which some researchers argue may cloud the interpretation of results. This way, referred as the standardized response mean, corresponds closely to the method of calculating a paired *t*-test.

The second way is the same mean changes in scores but divided by the standard deviation of the scores recorded at the first occasion. This second way centers on mean change in scores relative to background or natural variability of scores inherent to the PRO measure in the population sampled, variability that is free from an intervention's effect and extraneous events.

As an example for a single-group trial, the responsiveness of the SEAR questionnaire for erectile dysfunction in a single intervention study with sildenafil, with 93 subjects, was based on an ES defined as the mean change in scores from baseline divided by the standard deviation score at baseline (Althof et al. 2003). The magnitude of the change was quite high for most aspects of the SEAR questionnaire [Sexual Relationship Satisfaction, ES = 1.6; Confidence, ES = 1.0; Self-Esteem, ES = 1.1] and moderate for one (Overall Relationship Satisfaction, ES = 0.6), suggesting that the SEAR questionnaire is responsive for detecting psychosocial gains from a known beneficial intervention.

These two ways of computing an ES for a single group can be modified when comparing two interventions from two independent groups of participants. For two independent groups, the numerator can be the difference in means between two

independent groups, and the denominator can be the corresponding pooled standard deviations of scores from the two groups [which is how Cohen (1988) defined effect size and its magnitude of 0.2 as "small," 0.5 as "medium," and 0.8 as "large" effects]. Alternatively, the denominator can be the baseline standard deviation of scores pooled from the two groups, which some researchers may found preferable, where the magnitude of the mean difference between interventions is relative to the normal variability of measurement (before intervention).

As an example from a two-group trial, in double-blind placebo-controlled study of sildenafil, which enrolled 256 subjects, sizable treatment differences were observed and effect sizes on the SEAR questionnaire were calculated as the difference in the mean change scores between treatment groups divided by the pool standard deviation of scores at baseline (O'Leary et al. 2006). Large effect sizes of the differences in mean changes were obtained between active treatment and placebo treatment for the Self-Esteem subscale (ES = 0.84), for the Sexual Relationship Satisfaction domain (ES = 1.02), for the Confidence subscale (ES = 0.86); a moderate-to-large effect size was found for the Overall Relationship Satisfaction domain (ES = 0.63).

5.3.2 Probability of Relative Benefit

Differences between treatment groups at a specific follow-up time or change from baseline can be evaluated nonparametrically with the Wilcoxon rank-sum test using ridit analysis (Acion et al. 2006). This type of analysis is well-suited for ordinal responses at the item level or subscale or total scale levels. The Mann-Whitney rank-sum U statistic from the Wilcoxon rank-sum test gets converted, using ridit analysis, to a probability that represents the chance that a randomly selected patient from the treatment group has a more favorable response than a randomly selected patient from the control group. For instance, the method addresses the question, what is the likelihood that a randomly selected patient in the treatment group would have greater reduction in pain relative to a randomly selected patient in the control group?

As an illustration based on the literature, consider again the Self-Esteem And Relationshp (SEAR) questionnaire for men with erectile dysfunction. Here data were combined from two 12-week, double-blind, placebo-controlled, flexible-dose sildenafil trials having identical protocols: one conducted in the United States and the other in Mexico, Brazil, Australia, and Japan (Cappelleri et al. 2008). Response categories of each SEAR item used a 4-week reference period and were based on a five-point scale (1 = almost never/never, 2 = a few times, 3 = sometimes, 4 = most times, 5 = almost always/always). The difference (sildenafil versus placebo) in the change from baseline to week 12 was evaluated with the Wilcoxon rank-sum test using ridit analysis.

The probability of increased psychosocial benefit from baseline to week 12 was higher with sildenafil for each SEAR item (two-sided $P < 0.001$) and ranged from 0.60 ("My partner was unhappy with the quality of our sexual relations" [item reverse-scored]) to 0.72 ("I was satisfied with my sexual performance"). Across all items, the

average probability was 0.67 (standard deviation of 0.04) that a randomly selected patient in the sildenafil group would have a more favorable psychosocial change relative to a randomly selected patient in the placebo group.

5.3.3 *Cumulative Distribution Functions*

As a graphical representation to array all possible responder cutoffs, cumulative distribution function can display a continuous plot of the observed change (or percent change) from baseline on the horizontal axis and the cumulative percent of patients experiencing up to that change on the vertical axis, which negates the need for a specific or single responder definition. In essence each of the possible change score has a turn as a responder cutoff. Consider a situation where lower change or more negative scores are better or more favorable (Fig. 5.5). In Fig. 5.5, 70% of the subjects in the experimental group had scores of 10 or less (that is, 10 or better) compared with 55% of the subjects in the control group. The consistent horizontal separation between the distribution functions suggests that the treatment was beneficial relative to control over the entire range of changes.

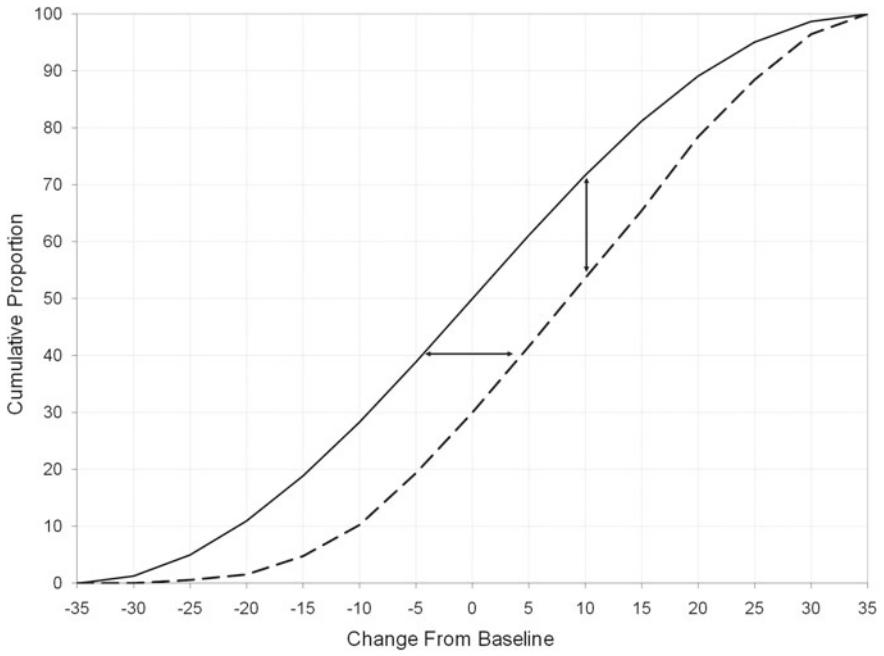


Fig. 5.5 Illustrative cumulative distribution functions of two treatments groups where more negative change scores are better (solid line, experimental group; dashed line, control group). Source: Cappelleri et al. (2013)

There are multiple ways to portray cumulative distribution plots. A key point here is that cumulative distribution plots be labeled and specified clearly to allow for easy and clear interpretation with respect to the directionality of score changes (e.g., whether positive changes indicate improvement or deterioration) and their associated cumulative percentages of patients. Such cumulative distribution response curves, one for each treatment group, would allow for a variety of response thresholds to be examined simultaneously and would encompass all available observed data.

Cumulative distribution plots are most compelling and best suited for interpretation when there is no or minimal overlap in the curves between treatments. When there is some or considerable overlap, the cumulative treatment curves interact and interpretation gets clouded. In such a case judgment is needed on what cutoff scores are considered clinically most plausible.

5.4 Mediation Models

A mediation model is one that seeks to identify and explain the mechanism that underlies an observed relationship between a predictor or independent variable (e.g., treatment group) and an outcome or dependent variable (e.g., sleep disturbance) via the inclusion of a third explanatory variable (e.g., pain), known as a mediator variable. Any of these three variables may be a PRO measure (e.g., sleep quality and pain may be PRO measures; an independent variable may also be a PRO measure). Mediation model is gaining currency in the application of PRO measures (Fairclough 2010; Cappelleri et al. 2013) and full-length monographs have been devoted to mediation analysis (e.g., Iacobucci 2008; MacKinnon 2008; VanderWeele 2015). In pharmaceutical studies, for example, mediation models can help to elucidate the mechanism of action of a drug or provide an understanding on the interrelationship of PRO measures to other variables, thereby advancing interpretation of PRO measures themselves.

Rather than hypothesizing a direct causal relationship between the predictor and the outcome, a mediation model postulates that the predictor variable not only affects the outcome variable directly, but also affects the mediator variable, which in turn also affects the outcome variable. The mediator variable, therefore, serves to clarify the nature of the relationship between predictor and outcome variables. The postulated underpinning for a mediation model is driven by the theoretical or conceptual framework and the research objective.

It should be emphasized that no technique, including mediation analysis and other forms of structural equation models, can definitely prove causation. Rather, the purpose of mediation analysis (and such path analyses in general) is to determine whether the hypothesized causal inferences by a researcher are harmonious with the data. If the mediation model does not fit the data, then revisions are needed because then one or more of the model or content-based assumptions are not correct or need to refined. If the mediation model is consistent with the data, this does not prove causation. Instead, it shows that the assumptions made are not contradicted and *may be valid*. It only *may be valid* because other models and assumptions may also fit

the data. Making causal inferences between variables is tricky business and a serious subject. The extent that one variable may cause another depends on the research design, including in part the temporal sequence of the variables and the plausibility of relations as informed by knowledge of the subject matter.

The approaches discussed in this section assume that the mediator and outcome are continuous variables (or variables taken to be continuous). The predictor or mediator or outcome may be a categorical (e.g., binary) variable, as well as continuous. Mediation models are given detailed exposition, including for categorical variables, elsewhere (Iacobucci 2008; MacKinnon 2008; VanderWeele 2015).

In this section, the basic elements of the single mediator model are described, the basic model is formulated, and then a real-life application using PRO measures is given.

5.4.1 Basic Elements

Research has often focused on the relation between two variables, say, X and Y. Such research includes situations where the explanatory (predictor) variable X can be considered a possible cause of the outcome variable Y, as when, for example, subjects are randomized to interventions of the treatment group variable X.

A theoretical premise may posit that an intervening (mediator) variable is an indicative measure of the process through which a predictor is thought to affect an outcome. The objective is to assess the extent to which the effect of the predictor variable on the outcome variable is indirect via the mediator or, alternatively, is otherwise direct, which captures all other effects.

As diagrammed in Fig. 5.6, mediation in its simplest form is represented by a third variable (M, the mediator), so that the predictor X influences the mediator M which, in turn, influences the outcome Y (X affects M and then M affects Y). Therefore, a natural question becomes what fraction of the total effect of X on Y is the direct effect and what fraction of the total effect of X on Y is the indirect effect mediated through the mediator M. The direct effect represents all other possible effects other than those attributed to the mediator.

There are essentially four assumptions of mediation models: (1) no unmeasured confounding of the predictor-outcome relationship, (2) no unmeasured confounding of the predictor-mediator relationship, (3) no unmeasured confounding of the mediator-outcome relationship, and (4) no mediator-outcome confounder that is affected by the predictor (no interaction between the predictor and mediator on the outcome) (VanderWeele 2015). The first 2 of these assumptions are automatically satisfied if the treatments, as levels of the predictor variable, were randomized. For more detail about these assumptions, including how to assess them and to use sensitivity analysis to help assess how robust results are to violations in the assumptions, the reader is referred elsewhere (VanderWeele 2015). If an assumption is not met, interpretation of results and manifestation of conclusions should be qualified appro-

priately, such as limiting interpretation and ensuing conclusions to merely association rather than causation.

5.4.2 Basic Model

The mediation model portrayed by Fig. 5.6 for the subject j can be denoted by the following equations

$$Y_j = i_1 + b \times X_j + c \times M_j + e_{1j} \tag{5.1}$$

$$M_j = i_2 + a \times X_j + e_{2j} \tag{5.2}$$

where

Y_j and M_j	are the outcomes for subject *j;*
i_1 and i_2	are the overall intercepts;
a	is the overall slope in Eq. 5.2, representing effect of the independent variable X on the mediator variable M;
b	is the overall slope in Eq. 5.1, representing direct effect of the independent variable X on the variable Y*;*
c	is the overall slope in Eq. 5.1, representing effect of the mediator variable M on the variable Y; and

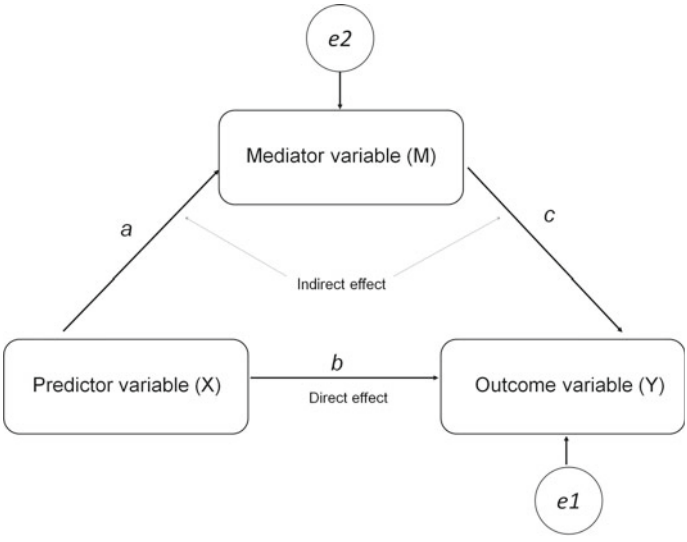


Fig. 5.6 Basic mediation model: the predictor X influences the outcome variable Y directly and via the mediator (M). Source: Cappelleri et al. (2013)

e_{1j} and e_{2j} are the error terms [assumed to be from normal distribution with mean *0* and variance σ_1^2 and σ_2^2, that is, $e_{1j} \sim N(0, \sigma_1^2)$ and $e_{2j} \sim N(0, \sigma_2^2)$].

It should be emphasized that it is assumed that the mediation model is correctly specified to answer the research question of interest from a well-defined hypothesis in order to draw reliable inferences. Under this hypothesis framework, the data provide empirical evidence regarding the postulated inter-relationship among variables. If the hypothesis changes to include or exclude certain variables, with possible changes in the linkages between variables, then the model formulation would need to be modified accordingly. Hence the model formulation should be fully aligned to its hypothesized framework.

Replacing M_j in Eq. 5.1 by M_j from Eq. 5.2 allows Y_j to be represented as

$$\begin{aligned} Y_j &= i_1 + (b \times X_j) + c \times [i_2 + (a \times X_j) + e_{2j}] + e_{1j} \\ Y_j &= [i_1 + (c \times i_2)] + \{[b + (c \times a)] \times X_j\} + [(c \times e_{2j}) + e_{1j}]. \end{aligned} \tag{5.3}$$

Equation 5.3 can be considered as the representation of the total effect of the variable X on variable Y, after accounting for the presence of the mediator M. The first part $[i_1 + (c \times i_2)]$ is constant and represents the intercept, the second part $[b + (c \times a)]$ represents the slope of this total effect, and the third part $[(c \times e_{2j}) + e_{1j}]$ represents the error term. If variable X_j represents treatment with values of 0 for placebo and value of 1 for the active treatment, then $[b + (c \times a)]$ represents total effect of the drug on the outcome Y after accounting for placebo. Coefficient b represents the direct effect of X_j on variable Y_j. It is worthwhile to note that the term "direct effect" is somewhat misleading—this effect actually represents all other possible paths (excluding path through the mediator M) from the independent variable X to the outcome Y. And expression $(c \times a)$ represents the indirect effect of X on Y through the mediator M. The mediation modeling can be viewed as an attempt to decompose the total effect of X on Y to better understand mechanism of action of X or its inter-relationship to M and Y.

Now we are ready to answer the main question: What fraction of the total effect of X on Y is the direct effect and what fraction of the total effect of X on Y is the indirect effect mediated through the mediator M? The percentage of the total effect that is the direct effect ("the direct effect of X on Y") can be expressed as:

$$direct\ effect = 100\% \left(\frac{b}{b + (c \times a)} \right). \tag{5.4}$$

In Eq. 5.4 the fraction was multiplied by 100% to represent the effect as a percentage of the total effect of X on Y.

The percentage of the total effect that is an indirect effect of X on Y ("the indirect effect of X on Y") via the mediator M can be expressed as:

$$indirect\ effect = 100\% \left(\frac{c \times a}{b + (c \times a)} \right). \tag{5.5}$$

Complete (100%) mediation is the case in which the variable X no longer directly affects Y, so the path coefficient b is zero. On the other hand, no mediation occurs when the total effect of X on Y exists entirely through the direct effect, so that the coefficient b is non-zero and $(c \times a)$ is zero. Partial mediation is the case in which the direct path coefficient b and indirect path coefficient $(c \times a)$ are both non-zero.

5.4.3 Example

Consider one study in which 745 patients were randomized to placebo or three different doses of study medication over 14 weeks (Russell et al. 2009). Specifically, we consider the direct and indirect effects of treatment with pregabalin 300, 450, and 600 mg (each versus placebo) on patient-reported sleep disturbance (range: 0–100, where higher scores reflect more sleep disturbance) from the Medical Outcomes Study Sleep Scale. This outcome, with a one-week recall period in this study, was assessed at Week 14, the end of the study.

The mediator was patient-reported daily diary pain score, based on an 11-point numeric rating scale (0 = no pain to 10 = worst possible pain) in the past 24 h. Like sleep disturbance scores, pain scores were assessed and culminated at Week 14; pain scores were based on the average rating over the last 7 days of the study (Week 14). A set of simultaneous linear multiple regression equations was postulated to quantify treatment-related improvements in sleep disturbance that appeared to be due to reductions in pain (indirect treatment effect) and treatment-related improvements in sleep outcomes that were not explained, or mediated, by reductions in pain (direct treatment effect).

Figure 5.7 depicts the pathways.

The total effect of pregabalin 300 mg, 450 mg, and 600 mg relative to placebo on sleep disturbance scores was a mean reduction (improvement) of 9.9, 12.5, and 15.2, respectively. The mediation model showed that 80, 73, and 75.6% (all significantly different from zero, two-sided $p < 0.0001$) of the reduction in sleep disturbance were direct effects of the treatments (respectively, pregabalin 300, 450, and 600 mg; all $p < 0.0001$) themselves, while the remaining 20% (not significant) for 300 mg, 27% ($p = 0.0153$) for 450 mg, and 24.4% ($p < 0.0027$) for 600 mg were mediated via pain. Under the assumption that model and content-based assumptions are met, the direct effect of study mediation on sleep disturbance reflects the effect of study medication independent of changes in pain, while the indirect effect of the medicine on sleep disturbance represents the part mediated via pain (prompted by the analgesic effects of the medicine).

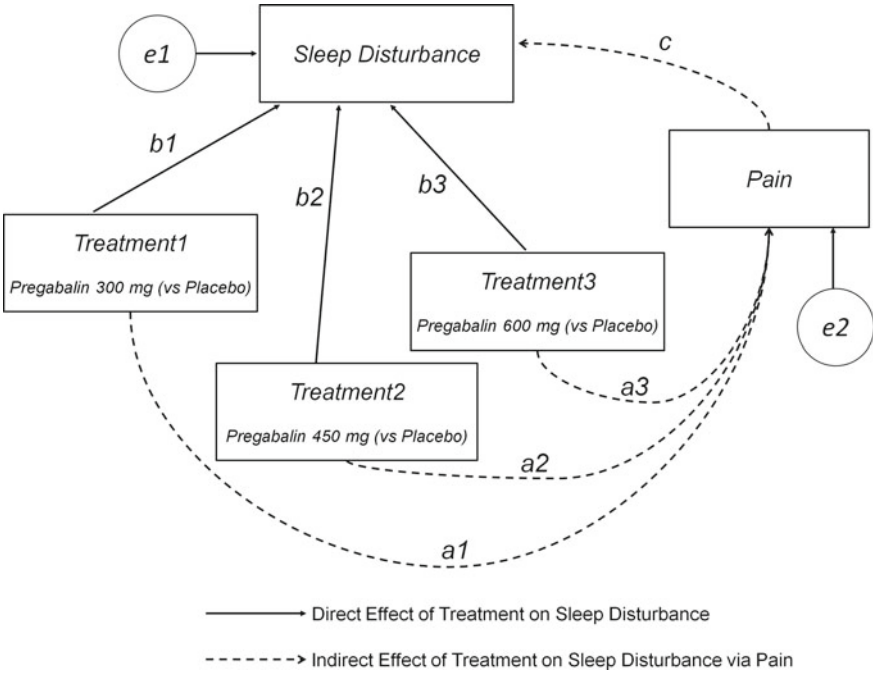


Fig. 5.7 Direct and indirect effects of pregabalin on sleep disturbance with pain as the mediator. Source: Cappelleri et al. (2013)

5.5 Summary

Useful interpretation of score values or score changes on patient-reported outcomes can be valuable in designing studies, evaluating interventions, educating consumers, and informing health-policy makers involved with regulatory, reimbursement, and advisory agencies. Unlike certain objectives outcomes like blood pressure, subjective outcomes often lack the historical, empirical, and clinical thread to draw from for meaningful interpretation.

This chapter focuses on enriching or advancing the interpretation of patient-reported outcomes, a topic central to and commensurate with their impact. The logic and rationale of two broad methods—anchor-based and distribution-based—are elucidated. Five anchor-based approaches are highlighted: percentages based on thresholds, criterion-group interpretation, content-based interpretation, clinically important difference, and clinically important responder. Three distributed-based approaches are described: effect size, probability of relative benefit, and cumulative distribution functions.

Receiving less attention, a third approach to enhance interpretation of patient-reported outcomes—mediation models—is also described. In its simplest form, mediation analysis enables the total effect of a predictor variable on an outcome

variable to be partitioned into an indirect effect via a mediator variable and a direct effect attributed to everything else. The formulation of and results from a mediation model depend on the assumption made and postulated framework posed by the research objective. Throughout the chapter, illustrative and real-life applications are provided to complement and supplement the exposition.

Acknowledgements This chapter draws directly from material in Chaps. 9 and 11 of our monograph, Cappelleri et al. (2013). *Patient-reported outcomes: Measurement, implementation and interpretation*. Boca Raton, Florida: Chapman & Hall/CRC Press.

References

Acion, L., Peterson, J. L., Temple, S., & Arndt, S. (2006). Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine, 25,* 591–602.

Althof, S. E., Cappelleri, J. C., Shpilsky, A., Stecher, V., Diuguid, C., Sweeney, M., et al. (2003). Treatment responsiveness of the Self-esteem and Relationship (SEAR) questionnaire in erectile dysfunction. *Urology, 61,* 888–893.

Bennett, R. M., Bushmakin, A. G., Cappelleri, J. C., Zlateva, G., & Sadosky, A. B. (2009). Minimally clinically important difference in the Fibromyalgia Impact Questionnaire (FIQ). *Journal of Rheumatology, 36,* 1304–1311.

Cappelleri, J. C., Rosen, R. C., Smith, M. D., Quirk, F., Maytom, M. C., Mishra, A., et al. (1999). Some developments on the International Index of Erectile Function (IIEF). *Drug Information Journal, 33,* 179–190.

Cappelleri, J. C., Althof, S. E., O'Leary, M. P., & Tseng, L. J. (2008). On behalf of the US and International SEAR Study Group: Analysis of single items on the Self-Esteem And Relationship questionnaire in men treated with sildenafil citrate for erectile dysfunction: Results of two double-blind placebo-controlled trials. *BJU International, 101,* 861–866.

Cappelleri, J. C., Bushmakin, A. G., McDermott, A., Dukes, E., Sadosky, A., Petrie, C. D., et al. (2009). Measurement properties of the medical outcomes study sleep scale in patients with fibromyalgia. *Sleep Medicine, 10,* 766–770.

Cappelleri, J. C., Zou, K. H., Bushmakin, A. G., Alvir, J. M. J., Alemayehu, D., & Symonds, T. (2013). *Patient-reported outcomes: Measurement, implementation and interpretation*. Boca Raton, Florida: Chapman & Hall/CRC Press.

Chang, C. H., & Reeve, B. B. (2005). Item response theory and its applications to patient-reported outcomes measurement. *Evaluation in the Health Professions, 28,* 264–282.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum.

Coon, C. D., & Cappelleri, J. C. (2016). Interpreting change in scores on patient-reported outcome instruments. *Therapeutic Innovation & Regulatory Science, 50,* 22–29.

Copay, A. G., Subach, B. R., Glassman, S. D., Polly, D. W., Jr., & Schuler, T. C. (2007). Understanding the minimum clinically important difference: A review of concepts and methods. *The Spine Journal, 7,* 541–546.

Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology, 56,* 395–340.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. New York, NY: Cambridge University Press.

Fairclough, D. L. (2010). *Design and analysis of quality of life studies in clinical trials* (2nd ed.). Boca Raton, Florida: Chapman & Hall/CRC Press.

Fayers, F. M., & Machin, C. (2016). *Quality of life: The assessment, analysis and interpretation of patient-reported outcomes* (3rd ed.). Chichester, England: Wiley.

Food and Drug Administration (FDA): Guidance for Industry. (2009). Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. December 2009. Retrieved December 29, 2016. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf.

Guyatt, G. H., Osoba, D., Wu, A., Wyrwich, K. W., & Norman, G. R. (2003). The Clinical Significance Consensus Meeting Group: Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings, 77,* 371–383.

Iacobucci, D. (2008). *Mediation analysis*. Thousand Oaks, California: SAGE Publications.

Hays, R. D., Farivar, S. S., & Liu, H. (2005). Approaches and recommendations for estimating minimally important difference for health-related quality of life measures. *Journal of Chronic Obstructive Pulmonary Disease, 2,* 63–67.

King, M. T. (2011). A point of minimal important difference (MID): A critique of terminology and methods. *Expert Reviews in Pharmacoeconomics & Outcomes Research, 11,* 171–184.

Kleinbaum, D., & Klein, M. (2010). *Logistic Regression* (3rd ed.). New York, NY: Springer.

Lewis, J. A. (2004). In defence of the dichotomy. *Pharmaceutical Statistics, 3,* 77–79.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. LLC, New York, NY: Taylor & Francis Group.

Mamolo, C. M., Bushmakin, A. G., & Cappelleri, J. C. (2015). Application of the Itch Severity Score in patients with moderate-to-severe plaque psoriasis: Clinically important difference and responder analysis. *Journal of Dermatological Treatment, 26,* 121–123.

Marquis, P., Chassany, O., & Abetz, L. (2004). A comprehensive strategy for the interpretation of quality-of-life data based on existing methods. *Value in Health, 7,* 93–104.

McLeod, L. D., Coon, C. D., Martin, S. A., Fehnel, S. E., & Hays, R. D. (2011). Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert Reviews of Pharmacoeconomics & Outcomes Research, 11,* 163–169.

McLeod, L. D., Cappelleri, J. C., & Hays, R. D. (2016). Best (but of forgotten) practices: Expressing and interpreting meaning and effect sizes in clinical outcome assessments. *American Journal of Clinical Nutrition, 103,* 685–693 (with erratum).

McLeod, L. D., Fehnel, S. E., & Cappelleri, J. C. (2018). Patient-peported outcome measures: Development and psychometric validation. In: K. E. Peace, D-G. Chen, & S. Menon (Eds.), *Biopharmaceutical applied statistics symposium: Design of clinical trials*. Vol. 1, Chapter 13. Singapore: Springer Nature Singapore Pte Ltd.

O'Connell, A. A. (2005). *Logistic regression models for ordinal response variables*. Thousands Oaks, California: SAGE Publications.

O'Leary, M. P., Althof, S. E., Cappelleri, J. C., Crowley, A., Sherman, N., & Duttagupta, S. (2006). On behalf of the US SEAR Study Group. 2006: Self-esteem, confidence, and relationship satisfaction in men with erectile dysfunction treated with sildenafil citrate: A multicenter, randomized, parallel-group, double-blind, placebo-controlled study in the United States. *Journal of Urology, 175,* 1058–1062.

Revicki, D., Erickson, P. A., Sloan, J. A, Dueck, A., Guess, H., Santanello, N. C., & the Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group (2007). Interpreting and reporting results based on patient-reported outcomes. *Value in Health, 10,* S116–S124.

Russell, I. J., Crofford, L. J., Leon, T., Cappelleri, J. C., Bushmakin, A. G., Whalen, E., Barrett, J. A., & Sadosky, A. (2009). The effects of pregabalin on sleep disturbance symptoms among individuals with fibromyalgia syndrome. *Sleep Medicine, 10,* 604–610.

Snapinn, S. M., & Jiang, Q. (2007). Responder analysis and the assessment of a clinically relevant treatment effect. *Trials, 8,* 31. https://doi.org/10.1186/1745-6215-8-31.

Streiner, D. L., Norman, G. R. J., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). New York, NY: Oxford University Press.

Thompson, J. R., Cappelleri, J. C., Getter, C., Pleil, A., Reichel, M., & Wolf, S. (2007). Enhanced interpretation of instrument scales using the Rasch model. *Drug Information Journal, 41,* 541–550.

VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. New York, NY: Oxford University Press.

Ware, J. E., Kosinski, M., Bjorner, J. B., Turner-Bowker, D. M., Gandek, B., & Maruish, M. E. (2007). *User's manual for the SF-36v2® health survey* (2nd ed.). Lincoln, Rhode Island: Quality Metric Incorporated.

Chapter 6
Network Meta-analysis

Joseph C. Cappelleri and William L. Baker

6.1 Introduction

A systematic literature review encompasses an explicit and detailed description of how a review on a topic was conducted. Systematic reviews of randomized controlled trials (RCTs) are considered the standard basis for evidence-based health-care decision-making for clinical treatment guidelines and reimbursement policies. Many systematic reviews use meta-analysis to combine quantitative results of similar and comparable studies in summarizing the available evidence. "Meta-analysis" may be defined as the statistical analysis of data from multiple studies. A meta-analysis typically identifies data systematically, summarizes results, and evaluates quantitatively sources of heterogeneity and bias (Borenstein et al. 2009; Cappelleri et al. 2010).

Meta-analysis offers several benefits. It may be used to address uncertainty and heterogeneity when results of studies disagree, to increase statistical power for primary outcomes and subgroups, to improve estimates of treatment effect, and to lead to new knowledge and formulation of new questions. On the other hand, meta-analysis is based only on what information and studies are available, possible resulting in publication bias and the "apples and oranges" phenomenon of mixing different studies that may compromise the quality and generalizability of results. To address such criticisms, the researcher should prepare a protocol that includes well-defined criteria and objectives for including the studies in a meta-analysis, as well as plans for subgroup analyses and regression analyses that may examine differences (heterogeneity) in treatment effect among studies. Another way to conduct a credible meta-analysis

J. C. Cappelleri (✉)
Pfizer Inc, 445 Eastern Point Road, MS 8260-2502, Groton, CT 06340, USA
e-mail: joseph.c.cappelleri@pfizer.com

W. L. Baker
University of Connecticut, Storrs, CT, USA
e-mail: william.baker_jr@uconn.edu

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_6

is to adhere to guidelines on evaluating and reporting systematic reviews and meta-analyses (Liberati et al. 2009).

A traditional pairwise meta-analysis of RCTs typically involves a direct (head-to-head) comparison of effects between two treatments across trials. In the absence of direct head-to-head evidence between two treatments of interest, an indirect comparison can provide useful evidence for the difference in treatment effects between competing interventions, which otherwise would be wanting, and for judiciously selecting the best choice(s) of treatment (Snedecor et al. 2014). For example, if two particular treatments have never been compared against each other, head-to-head, but these two treatments have been compared against a common comparator, then an indirect treatment comparison can use the relative effects of the two treatments versus the common comparator. Even when some direct evidence exists, it would be useful to combine it with results from indirect evidence on the same pair of treatments, which may make the assessment more precise and informed.

Based on indirect evidence, with or without direct evidence, network meta-analysis (NMA) can be conducted if both treatments have been compared to a common comparator. The estimate of treatment effect obtained from such an analysis is referred to as "indirect evidence." That is, an indirect estimate of the effect of treatment A over B can be obtained by comparing trials of A versus C and B versus C. Extending this concept, NMAs can also allow simultaneous comparison of more than two treatments. Thus, in its broadest sense, NMA can be defined as a statistical combination of all available evidence for an outcome from several studies across multiple treatments to generate estimates of pairwise comparisons of each intervention to every other intervention within a network (Caldwell et al. 2005; Jansen et al. 2011; Lu and Ades 2004).

In this chapter we provide an introduction to network meta-analysis of randomized controlled trials on study-level or aggregate data on the same disease of interest. In doing we describe evidence networks; the analytic methodology with fixed-effect and random-effects models, including from a Bayesian perspective, along with an application; the assumptions of homogeneity, similarity, and consistency; and a few special topics on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) guidance, individual patient data, and population-based adjusted indirect comparisons.

6.2 Evidence Networks

Network meta-analyses are so named because all of the treatments analyzed are connected to every other treatment via a network of randomized comparisons, sometimes referred to as the "evidence network." In the evidence network, each treatment is depicted as a node and the RCTs containing the treatments are represented as lines connecting the nodes. There may be multiple RCTs involving the same pair of treatments.

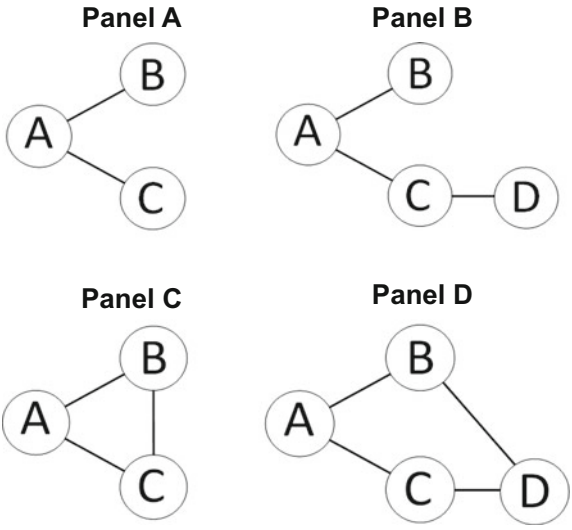


Fig. 6.1 Connected networks of randomized controlled trials. *Source* Jansen et al. Value in Health 17, 157–173 (2014)

An example of a disconnected network would be a network with an AB trial and a CD trial. Because AB and CD do not share a common intervention, any indirect assessment (e.g., BD or AC) is fraught with danger and a substantial risk of bias. Figure 6.1, on the other hand, shows examples of connected networks of RCTs, with varying levels of complexity. A connected network means that any two treatments can be compared indirectly through (one or more) intermediate common comparators or directly (head to head). In Fig. 6.1, the nodes represent interventions and the edges (or connections) imply that one or more RCTs have included their respective treatments directly.

Figure 6.1a includes AB studies (treatment A compared directly with treatment B) and AC studies (treatment A compared directly with treatment C) that allow the relative effect of B–C to be obtained indirectly. The network in Fig. 6.1b adds CD studies and, in doing so, engenders additional indirect comparisons: AD (through C), BC (through A), and BD (through BC and CD). In Fig. 6.1c, treatment pairs AB, AC, and BC are each compared both directly and indirectly. Figure 6.1d contains direct and indirect evidence for all pairwise comparisons with the exception of AD and BC, for which there is only indirect evidence.

To convey more information on the available clinical trials, the size of each node can be made proportional to the number of patients receiving that treatment and thickness of the lines between treatments can be proportional to the number of available RCTs informing the comparison. An analysis combining both types of data (direct and indirect) are often referred to as mixed treatment comparisons and their joint inclusion can help strengthen the precision of treatment effects between a pair of treatments in the network.

The terms network meta-analysis, indirect comparisons, and mixed treatment comparisons are often used interchangeably. Technically, network meta-analysis is

a broader concept and can be used whenever the evidence base consists of two or more trials connecting three or more treatments. Indirect treatment comparisons and mixed treatment comparisons can be considered as sub-classifications within network meta-analyses.

6.3 Methodology and Application

6.3.1 Fixed-Effect Model

Examples of relative treatment effect for two groups include risk ratio, odds ratio and risk difference for binary outcomes; the natural logarithm of risk ratio and odds ratio are typically used to address the normality assumption and thereby to strengthen statistical inference before they are converted to their original metric for interpretation and final results. Corresponding examples of treatment effect for continuous outcomes include the mean difference and standardized mean difference (Borenstein et al. 2009).

With a fixed-effect model, it is assumed that there is no variation in the true (population) relative treatment effects across studies for a particular pairwise comparison; each estimate of treatment effect in each study is measuring the same underlying treatment effect common to all studies, with inferences pertaining to the same fixed set of studies. Observed differences for a particular comparison across studies occur solely due to chance (Borenstein et al. 2009).

When the evidence network consists of multiple pairwise comparisons (i.e., AB trials, AC trials, BC trials, and so on), the set of comparators usually varies among studies, complicating the notation. One approach labels the treatments A, B, C, and so on, and uses A as the primary reference treatment in the analysis. For each study, the approach then designates one treatment, b, as the base treatment. The labels can be assigned to treatments in the network in such a way that the base treatments follow A (i.e., B, C, and so on) and the non-base treatments in turn follow all the base treatments in the alphabet. In the various models, "after" refers to this alphabetical ordering. The general frequentist fixed effect model for network meta-analysis can then be specified as follows (Hoaglin et al. 2011):

$$\eta_{jk} = \begin{cases} \mu_{jb} & b = A, B, C, \quad if\ k = b \\ \mu_{jb} + d_{bk} = \mu_{jb} + d_{Ak} - d_{Ab} & k = B, C, D, \quad if\ k\ is\ \text{“after”}\ b \end{cases}$$

where

η_{jk} reflects the underlying outcome for treatment k in study j,
μ_{jb} is the outcome for treatment b in study j, and
d_{bk} is the fixed effect of treatment k relative to treatment b.

The d_{bk} are identified by expressing them in terms of effects relative to treatment A : $d_{bk} = d_{Ak} - d_{Ab}$, with $d_{AA} = 0$. For the underlying effects, this relation is a statement that explicitly assumes consistency between the direct effect and the indirect effect: the "direct" effect d_{bk} and the "indirect" effect $d_{Ak} - d_{Ab}$ are equal.

The corresponding general Bayesian fixed effect model would place a prior distribution on d_{Ak}. Bayesian methods combine the likelihood (roughly, the probability of the data as a function of the parameters) with a prior probability distribution (which reflects prior belief about possible values of those parameters) to obtain a posterior probability distribution of the parameters (Hoaglin et al. 2011; Sutton and Abrams 2001), which also holds for a random-effects model (next subsection, Sect. 6.3.2). The posterior probabilities provide a straightforward way to make predictions, and the prior distribution can incorporate various sources of uncertainty. For parameters such as treatment effects, the customary prior distributions are non-informative. The assumption that, before seeing the data, all values of the parameter are equally likely minimizes the influence of the prior distribution on the posterior results. However, when information on the parameter is available (e.g., from observational studies or from a previous analysis), the prior distribution provides a natural way to incorporate it.

6.3.2 *Random-Effects Model*

If there is heterogeneity and, therefore, variation across trials in true (or underlying) relative treatment effects for a particular pairwise comparison, random effects models are appropriate. A random-effects model approach assumes that the trial-specific treatment effect can be described as each having its own normal distribution, which can differ from different studies (unlike the fixed-effect case), and whose estimate constitutes a point of an overall normal distribution whose central position represents the combined effect and whose standard deviation reflects the heterogeneity of treatment effects (Borenstein et al. 2009). With a random-effects model for a network meta-analysis, the variance reflecting heterogeneity is often assumed to be constant for all pairwise comparisons.

As an extension of the frequentist fixed-effect model, the frequentist random-effects model replaces d_{bk} with δ_{jbk}, the trial-specific effect of treatment k relative to treatment b (Hoaglin et al. 2011). These trial-specific effects are drawn from a random-effects distribution: $\delta_{jbk} \sim N\left(d_{bk}, \sigma^2\right)$. Again, values of d_{bk} are identified by expressing them in terms of the primary reference treatment, A (again with $d_{AA} = 0$). As noted in the prior paragraph, this model typically assumes the same random-effect variance σ^2 for all treatment comparisons, but the constraint can be relaxed. (A fixed-effect model results if $\sigma^2 = 0$.) Hence:

$$\eta_{jk} = \begin{cases} \mu_{jb} & b = A, B, C, \quad if\ k = b \\ \mu_{jb} + \delta_{jbk} & k = B, C, D, \quad if\ k\ is\ \text{“after” } b \end{cases}$$

$$\delta_{jbk} \sim N\left(d_{bk}, \sigma^2\right) = N\left(d_{Ak} - d_{Ab}, \sigma^2\right)$$

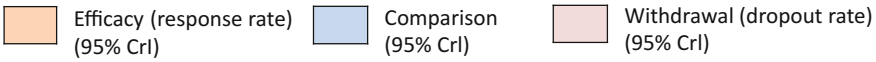
The corresponding random-effects Bayesian model would add a prior distribution on not only d_{Ak} (as is the case for the fixed-effect Bayesian model) but also σ.

6.3.3 Reporting and Interpreting

Results of random-effects models, as well as fixed-effect models, can be portrayed in a league table for each possible treatment comparison. Frequently, each cell of the table can contain the estimated treatment effect for an outcome and its 95% credible interval (if a Bayesian approach is taken) or 95% confidence interval (if a frequentist approach is taken) to account for uncertainty. This set of treatment effect sizes can be also presented in a forest plot against a common comparator.

Figure 6.2 illustrates a league table with five treatments (four active agents and placebo) and two outcomes (efficacy and withdrawal). Results are the odds ratio in the column-defining treatment compared with the odds ratio in the row-defining treatment. For results on efficacy, found in the lower diagonal, odds ratios higher than 1 favor the column-defining treatment. For example, the odds of successfully responding on an efficacy outcome for Drug C is 1.37 times that of Drug D (95% credible interval, 1.19–1.65). For results on withdrawal, found in the upper diagonal, odds ratios lower than 1 favor the column-defining treatment. For instance, the odds of withdrawing on placebo is 0.75 times (three-quarters) that of Drug B (95% credible interval, 0.55–0.95).

In the Bayesian framework, probabilities regarding the distribution of parameters can be calculated. In each Markov chain Monte Carlo cycle, each treatment k is ranked according the estimated effect size. Then the proportion of the cycles in which a given treatment ranks first out of the total gives the probability $P(k=1)$ that treatment k ranks first as the best among the available treatment options. Similar probabilities can be calculated for the being second best, third best, and so forth. All of these probabilities sum to one for each treatment and each rank. For each treatment, rank probabilities can be plotted against the possible ranks for a given treatment, resulting in "rankograms" (Salanti et al. 2011). In addition, cumulative ranking for each treatment enables the ranking of each treatment overall, thereby indicating which treatment is best overall, second best, and so forth. It is important to emphasize the effect sizes (from the league table or forest plot) over the ranking, because a good rank does not necessarily imply a large or clinically important effect size.



PLACEBO	1.01 (0.78-1.28)	0.75 (0.55-0.95)	1.06 (0.86-1.32)	0.87 (0.74-0.98)
0.99 (0.79-1.24)	**DRUG A**	0.74 (0.52-0.99)	1.07 (0.86-1.31)	0.90 (0.73-1.09)
1.09 (0.83-1.43)	1.15 (0.90-1.47)	**DRUG B**	1.42 (1.09-1.85)	1.09 (0.91-1.27)
0.80 (0.65-0.98)	0.82 (0.71-1.00)	0.75 (0.60-0.93)	**DRUG C**	0.80 (0.72-0.93)
1.08 (0.92-1.31)	1.10 (0.86-1.47)	0.99 (0.69-1.34)	1.37 (1.19-1.65)	**DRUG D**

Fig. 6.2 Illustrative league table on odds ratios for five treatments and two outcomes. CrI = credibility interval

6.3.4 Application

As an example, a random-effects network meta-analysis was implemented within a Bayesian framework using Markov chain Montel Carlo methods in WinBUGS (MRC Biostatistics Unit, Cambridge, UK) and applied to assess the effects of 12 new-generation antidepressants on major depression (Ciprani et al. 2009). Based on a systematic review of 117 randomized controlled trials (25,928 participants), antidepressants were quantified, compared, and ranked with respect to proportion of patients who responded to allocated treatment (efficacy) and, separately, the proportion who dropped out of the allocated treatment (acceptability). Clinically important differences existed between commonly prescribed antidepressants for both efficacy and acceptability.

6.4 Assumptions

Network meta-analyses combine data of multiple interventions across several RCTs to synthesize estimates of relative treatment effects to generate pairwise comparisons. The validity and accuracy of estimates from NMAs depend on the requirement that trials in the network are sufficiently comparable and similar to yield meaningful unbiased estimates. To that end, three assumptions underlie NMA methodology and should always be tested when possible: homogeneity, similarity, consistency (Jansen et al. 2011, 2014; Hoaglin et al. 2011; Donegan et al. 2013).

6.4.1 Homogeneity

Homogeneity, the first assumption, assumes that there is no significant variation (or if present, it is due to random chance) in treatment effects among studies of the same comparison. In other words, for example, are all AB trials (and, separately and independently, all AC trials) "comparable" and estimating the same treatment effect? This assumption is applicable to network meta-analyses as it is in pairwise meta-analyses. Homogeneity can be separately assessed for each collection of identical comparisons within the network using standard statistical measures, such as Q statistic or I^2 (or both) (Borenstein et al. 2009). If heterogeneity exists, then the possible sources should be explored and implementation of random-effects modeling, sensitivity analyses, subgroup analyses, or meta-regression should be considered if sufficient data are available.

6.4.2 Similarity

Trial evidence may be homogeneous within certain pairwise comparisons, but significant variation in trial characteristics across different comparisons within a network can still lead to biased estimates. This leads to the second assumption—*similarity* (or, sometimes referred to as *transitivity)*—that requires all trials included within a network to be "comparable" in terms of key factors that can be potential effect modifiers (such as patient baseline characteristics, trial design, outcome definition and/or measurement, and follow-up time) that may affect (relative) treatment effect. Here, different levels of an effect modifier may modify or differentially affect the treatment effect for a given pair of treatment, with the treatment effect depending on the level of the effect modifier.

Similarity—which cannot be formally tested and verified—can be gauged (though not proven) through quantitative techniques (sensitivity analysis, meta-regression, subgroup analysis) and assessed qualitatively using summary tables documenting relevant baseline characteristics of patients and description of studies. Therefore, substantial (or systematic) differences in effect modifiers can be judged by comparing study specific inclusion and exclusion criteria, baseline patient characteristics, and study characteristics that are expected to modify treatment effect.

The assumption of similarity is not violated if differences in baseline or study characteristics between trials do not modify or influence treatment effect. It is only when such characteristics are treatment effect modifiers that the estimated treatment effect becomes biased.

In RCTs the observed outcome with an intervention is the result of study characteristics, patient characteristics, and the treatment itself. In a placebo-controlled trial, the result of the placebo arm reflects the impact of study and patient characteristics on the outcome of interest, say outcome y, as shown in Fig. 6.3.

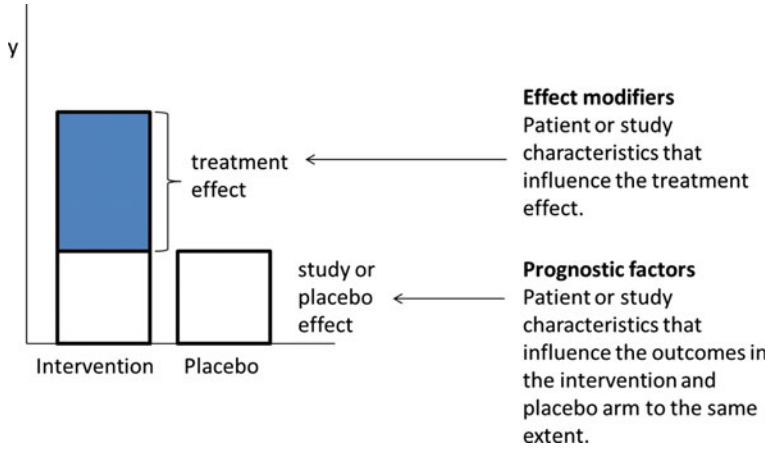


Fig. 6.3 Treatment effects, study effects, effect modifiers and prognostic factors in a randomized placebo-controlled trial. *Source* Jansen et al. Value in Health 17, 157–173 (2014)

In other words, the placebo response is the result of all known and unknown prognostic factors other than active treatment. We can call this the study effect. In the active intervention arm of the trial, the observed outcome y is a consequence of the study effect and a treatment effect. By randomly allocating patients to the intervention and placebo group, both known and unknown prognostic factors (as well as both measured and unmeasured prognostic factors) between the different groups within a trial are on average balanced. Hence, the study effect as observed in the placebo intervention arm is expected to be the same in the active intervention arm and, therefore, the difference between the active intervention arm and placebo intervention arm (say delta y) is attributable to the active intervention itself, resulting in a treatment effect (the blue box in Fig. 6.3).

Although a network meta-analysis is based on RCTs, randomization does not hold across the set of trials used for the analysis because patients are not randomized to different trials. As a result, there are situations where there are systematic differences in study characteristics or the distribution of patient characteristics across trials. In general, if there is an imbalance in the distribution of the effect modifiers across the different types of direct comparisons in a network meta-analysis, the corresponding indirect comparisons are biased.

Figure 6.4 provides an illustration where disease severity is known to be an effect modifier. Consider an indirect comparison of treatments B and C through an AB study (trial 1) and AC study (trial 2) where these two studies have different proportions of patients with moderate and severe disease. The indirect comparison of B versus C for the moderate disease population and, separately, the severe population are both valid. But the indirect comparison of B versus C for the overall population is biased because the distribution of the effect modifier severity is different for the AB and AC studies.

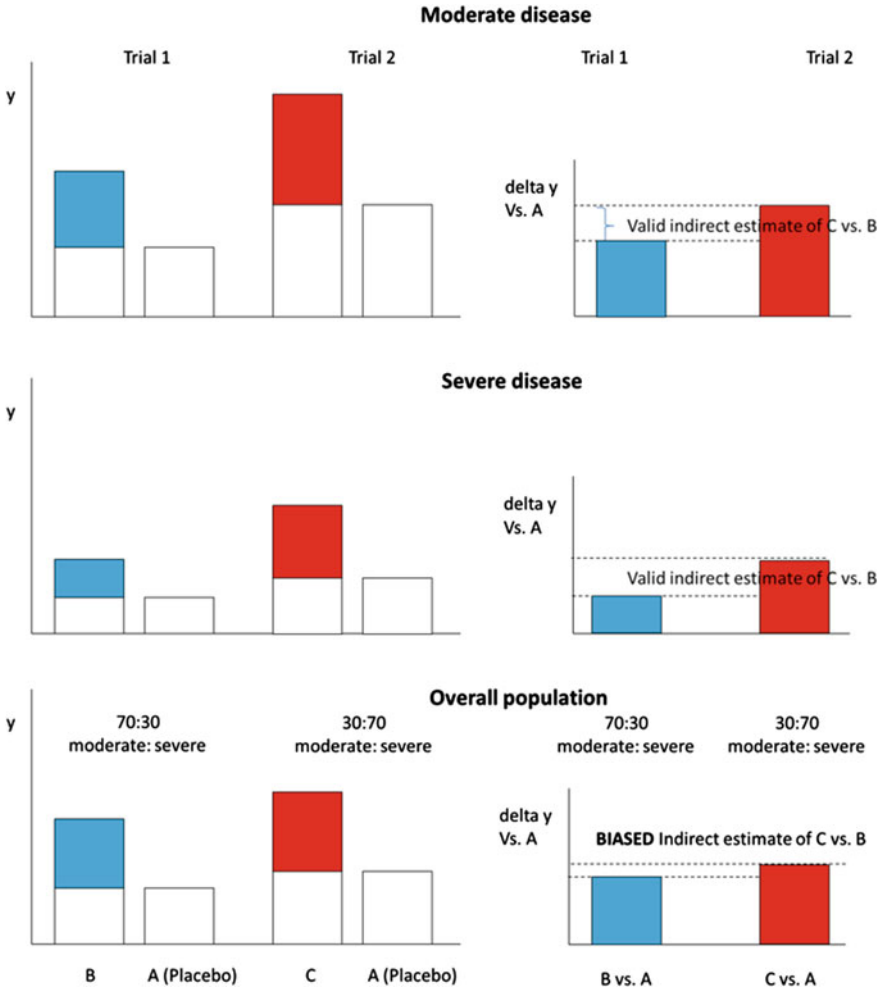


Fig. 6.4 Indirect comparisons with severity of disease known to be an effect modifier. *Source* Jansen et al. Value in Health 17, 157–173 (2014)

Suppose, on the other hand, that disease severity was *not* considered an effect modifier in a separate network meta-analysis. In this case, even if the set of proportions on disease severity was clearly different between AB and AC studies, as is the case in Fig. 6.4, then the indirect comparison of B versus C for the overall population would be unbiased instead of biased. It is important to acknowledge that there is always some risk of imbalances in unknown or unmeasured effect modifiers between studies evaluating different interventions. Accordingly, there is always a small risk of such residual confounding bias, even if all observed effect modifiers are balanced across the direct comparisons.

6.4.3 Consistency

When direct and indirect evidence are combined for a particular comparison, it is assumed that there is agreement between direct and indirect comparisons. This assumption is termed *consistency*, and it should be assessed in every NMA whenever possible. Figure 6.1c shows a simple closed loop network, where both direct and indirect evidence is possible for all pairwise comparisons. For example, an estimate of effect for B versus C can be obtained directly from the BC trial and can also be estimated indirectly from AC and AB trials. For this loop to be consistent, the direct estimate should be equivalent to the indirect estimate (i.e., $d_{BC} = d_{BA} - d_{CA}$). Of note, consistency is a property of closed loops of evidence and not individual comparisons. It is possible to state that AB, BC and AC comparisons are each separately consistent but stating that the AB comparison is consistent with the AC comparison has no meaning.

Inconsistencies can be caused by differences in treatment effect modifiers among the studies within a loop. Three independent studies forming a closed loop of evidence are unlikely to generate exact equality within a consistency evaluation. Published methods are available for evaluating consistency and its acceptable ranges (Dias et al. 2010, 2013).

6.5 Special Topics

6.5.1 PRISMA Guidance

Several guidances are available for conducting a proper NMA and for appraising a NMA (Ades et al. 2013; Jansen et al. 2014; Salanti et al. 2014; Hutton et al. 2015). For example, as an extension of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) for traditional pairwise treatment comparison (only two treatments given), a modified 32-item PRISMA extension checklist was developed to address what was deemed immediately relevant to the reporting of network meta-analyses (Hutton et al. 2015). Current PRISMA items were also clarified. This document presents the extension and provides examples of good reporting, as well as elaborations regarding the rationale for new checklist items and the modification of previously existing items from the original PRISMA statement.

Specifically, a checklist of items is included when reporting a systematic review of a NMA. Each checklist item pertains to particular section or topic that includes title, abstract, introduction, methods, results, discussion, and funding; in turn, each section or topic may have its subsections or subtopics. For instance, the Methods section includes a subsection on "risk of bias within individual studies" whose checklist item pertains to the description of methods for assessing risk of bias of individual studies and how this information is to be used in any data synthesis. Another subsection on Methods is "assessment of consistency" whose checklist item pertains to the

description of statistical methods used to evaluate the agreement of direct and indirect evidence in the treatment networks studies and to describe efforts taken to address its presence when found.

6.5.2 Individual Patient Data

Applications of study-level or aggregate meta-analysis cannot fully capitalize on the deeper information contained in the primary studies. A challenge in using aggregate data is that the association (or lack of association) between a patient-level covariate and (relative) treatment effects at the study level may be different from the true individual-level patient modification. A major benefit of meta-analysis with individual patient data is the ability to discern and quantify effect modification with more confidence and accuracy than can be performed with study-level meta-analysis. Individual patient data can be also performed within the context of NMA. Although it would add an additional layer of complexity to the analytical task, detailed formulation and direction are available on the theoretical and practical approaches relevant to the conduct of NMA with individual patient data (Veroniki et al. 2014). While existing methods appear suitable, more methodological advancements and improvements are welcomed.

Methods are also available for NMA of individual-level and aggregate-level data taken together (Jansen 2012). Non-linear network meta-analysis methods for combining both data types have been developed to reduce bias and uncertainty of direct and indirect treatment effects in the presence of heterogeneity. One method uses the same form for both types. Another method develops the model for aggregate data by integrating an underlying individual patient data model over the joint within-study distribution of covariates. This second method seems less affected by bias in situations with large interactions between treatment and patient-level covariates, probably at the cost of greater uncertainty. Having individual patient data available for a subset of studies can improve estimates of treatment effects in the presence of patient-level heterogeneity. Additional research remains.

6.5.3 Population-Adjusted Indirect Comparisons

Matched Adjusted Indirect Comparisons (MAICs) and Simulated Treatment Comparisons (STCs) have been increasingly applied in health technology assessment and, more specifically, to submissions to the National Institute for Health and Care Excellence (Phillippo et al. 2016). MAIC and STC are based, respectively, on propensity score reweighting and outcome regression (Ishak et al. 2015). Both are established methods of mapping a treatment effect observed in one population to an estimate of what would be observed in another, with a different distribution of prognostic factors and effect modifiers.

The novelty in MAIC and STC is first to apply these methods in the context of indirect comparisons, and second to rework them for a very specific scenario, in which a manufacturer has access to individual patient data from its own trial of product B against standard treatment A, but has access only to aggregate data on outcomes and covariates from a competitor AC trial.

MAIC and STC attempt to adjust for imbalances in baseline characteristics when forming an indirect comparison of treatments B and C. Standard methods for indirect comparisons assume that there is no imbalance in the distribution of effect modifiers in the AB and AC trials. Where effect modifiers are in imbalance, there is therefore a sound rationale for population-adjustment.

Nevertheless, based on a NICE technical support document on MAIC and STC that has been produced recently, a series of weaknesses in the MAIC and STC methods have been identified (Phillippo et al. 2016). Such limitations on these methods include the following: (1) they typically carry out indirect comparisons on the natural outcome scale, rather than on the usual linear predictor scale (i.e., log-odds scales for probabilities, log scale for rates), which raises questions about the interpretation of the model and of the indirect comparison; (2) they are often carried out in unconnected networks or with one-arm studies, called an "unanchored" or "unadjusted" indirect comparisons, with the degree of residual systematic error unknown; and (3) they can only deliver an estimate of the relative treatment effects in the AC population, which is very unlikely to be the target population for decision.

To address these concerns, recommendations for use of population-adjusted indirect comparisons have been rendered (Phillippo et al. 2016). These recommendations cover five areas: (1) the rationale for the use of population adjustment in NICE submissions, (2) justifying the use of population adjustment in both anchored and unanchored scenarios, (3) variables for which population adjustment is required, (4) generation of indirect comparison for the appropriate target population, and (5) reporting guidelines for analyses involving population adjustment. Moreover, research recommendations are given about the need for alternative approaches to population-adjusted indirect comparison and population-adjusted network meta-analysis intended to overcome the limitations of existing methods and, given this, the need for a plan on a comprehensive set of simulation studies, empirical studies, and illustrative applications of the new methods.

6.6 Summary

This chapter provides an introduction on NMA and targets its key concepts. In doing so, a descriptive framework is given on evidence networks; the analytic methodology with fixed-effect and random-effects models, including from a Bayesian perspective, along with an application; the assumptions of homogeneity, similarity, and consistency; and a few special topics on the PRISMA guidance, individual patient data, and population-based adjusted indirect comparisons.

Network meta-analysis can be considered an extension of traditional meta-analysis that includes multiple different pairwise comparisons across a range of different interventions to allow for multiple treatment comparisons in the absence of head-to-head evidence. Furthermore, the methodology can combine direct with indirect treatment comparisons, thereby synthesizing a greater share of the available evidence than traditional meta-analysis. Although the evidence networks underlying NMA typically include RCTs, randomization does not hold across trials and there is a risk of confounding bias, compromising internal validity. Accordingly, a NMA must be considered observational evidence, although a well-conducted NMA can provide a high quality assessment on the best available evidence.

References

Ades, A. E., Caldwell, D. M., Reken, S., Welton, N. J., Sutton, A. J., & Dias, S. (2013). Evidence synthesis for decision making 7: A reviewer's checklist. *Medical Decision Making, 33,* 679–691.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley.

Caldwell, D. M., Ades, A. E., & Higgins, J. P. T. (2005). Simultaneous comparison of multiple treatments: Combining direct and indirect evidence. *British Medical Journal, 331,* 897–900.

Cappelleri, J. C., Ioannidis, J. P. A., & Lau, J. (2010). Meta-analysis of therapeutic trials. In S. C. Chow (Ed.), *Encyclopedia of biopharmaceutical statistics, revised and expanded* (3rd ed., pp. 768–779). New York: Informa Healthcare.

Ciprani, A., Furukawa, T. A., Salanti, G., Geddes, J. R., Higgins, J. P. T., Churchill, R., et al. (2009). Comparative efficacy and acceptability of 12 new-generation antidepressants: A multiple-treatments meta-analysis. *Lancet*, *373*, 746–758.

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.

Dias, S., Welton, N. J., Sutton, A. J., & Ades, A. E. (2013). Evidence synthesis for decision making 4: Inconsistency in networks of evidence based on randomized controlled trials. *Medical Decision Making, 33,* 641–656.

Donegan, S., Williamson, P., D'Alessandro, U., & Smith, C. T. (2013). Assessing key assumptions of network meta-analysis. *Research Synthesis Methods, 4,* 291–323.

Hoaglin, D. C., Hawkins, N., Jansen, J. P., Scott, D. A., Itzler, R., Cappelleri, J. C., et al. (2011). Conducting indirect-treatment-comparison and network-meta-analysis studies: Report of the ISPOR task force on indirect treatment comparisons good research practices: Part 2. *Value in Health, 14,* 429–437.

Hutton, B., Salanti, G., Caldwell, D. M., Chaimani, A., Schmid, C. H., Cameron, C., et al. (2015). The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Annals of Internal Medicine, 162,* 777–784.

Ishak, K. J., Proskorovsky, I., & Benedict, A. (2015). Simulation and matching-based approaches for indirect comparisons of treatments. *Pharmacoeconomics, 33,* 537–549.

Jansen, J. P. (2012). Network meta-analysis of individual and aggregate level data. *Research Synthesis Methods*, *3*, 177–190.

Jansen, J. P., Fleurence, R., Devine, B., Itzler, R., Barrett, A., Hawkins, N., et al. (2011). Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: Report of the ISPOR task force on indirect treatment comparisons good research practices: Part 1. *Value in Health, 14,* 417–428.

Jansen, J. P., Trikalinos, T., Cappelleri, J. C., Daw, J., Andes, S., Eldessouki, R., et al. (2014). Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: An ISPOR-AMCP-NOC good practice task force report. *Value in Health, 17,* 157–173.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Annals of Internal Medicine, 151,* W65–W94.

Lu, G., & Ades, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine, 23,* 3105–3124.

Phillippo, D. M., Ades, A. E., Dias, S., Palmer, S., Abrams, K. R., & Welton, N. J. (2016). NICE DSU technical support document 18: Methods for population-adjusted indirect comparisons in submission to NICE. http://www.nicedu.org.uk. Accessed 1 May 2017.

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.

Salanti, G., Del Giovane, C., Chaimani, A., Caldwell, D. M., & Higgins, J. P. T. (2014). Evaluating the quality of evidence from a network meta-analysis. PLoS ONE *9*(7). https://doi.org/10.1371/journal.pone.0099682. Accessed 1 May 2017.

Snedecor, S. J., Patel, D. A., & Cappelleri, J. C. (2014). From pairwise to network meta-analysis. In G. Biondi-Zoccai (Ed.), *Network meta-analysis: Evidence synthesis with mixed treatment comparison* (pp. 21–41). New York: Nova Science Publishers.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research, 10,* 277–303.

Veroniki, A. A., Huedo-Medina, T. B., & Fountoulakis, K. (2014). Moving from study-level to patient-level data: Individual patient network meta-analysis. In G. Biondi-Zoccai (Ed.), *Network meta-analysis: Evidence synthesis with mixed treatment comparison* (pp. 223–244). New York: Nova Science Publishers.

Chapter 7
Detecting Safety Signals Among Adverse Events in Clinical Trials

Richard C. Zink

7.1 Introduction

Gaining a clear picture for the safety of any drug can be challenging, but when the necessary understanding of patient safety is at its greatest, sufficient insight into the tolerability of a treatment is often more difficult to attain. Establishing the effectiveness of a new therapy is often limited to a single primary outcome, with a handful of other secondary outcomes providing additional evidence of benefit. Safety outcomes, on the other hand, include the myriad of other data that are collected during the course of clinical development. Death and disease progression are obvious safety endpoints, and data for adverse events (AEs), laboratory abnormalities, vital signs, physical examinations, hospitalizations, electrocardiograms (ECGs), and patient-reported outcomes for quality-of-life can suggest other safety and tolerability concerns for the patient. Safety considerations can comprise efficacy outcomes as well, as there is the potential for these endpoints to worsen during the trial. The inherent multiplicity problem present when analyzing numerous endpoints is further complicated in several ways. First, and similar to efficacy endpoints, safety outcomes can be repeatedly measured over time. Second, safety outcomes have important characteristics to consider including duration, severity, and investigator's assessment of causal relationship to drug, resulting in numerous sensitivity analyses. Further, it is unclear which collection of event attributes would warrant consideration as the primary analysis. Third, many safety issues may occur spontaneously at any time during the trial, and often transpire between study visits. This adds complexity for summa-

R. C. Zink (✉)
TARGET PharmaSolutions, Chapel Hill, NC, USA
e-mail: rzink@targetpharmasolutions.com

R. C. Zink
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_7

rizing results across time, and may result in some level of missing data for events which, depending on the assumptions, could impact inference between study arms.

Though the previous paragraph suggests that multiplicity adjustments should be applied to limit the potential for type I errors, the analyst should proceed cautiously. Recall that most clinical trials are designed to establish efficacy for one or perhaps a small number of primary endpoints. Given the rarity of many safety outcomes, however, the sample sizes that are appropriate for efficacy endpoints, even when combined across several trials, result in treatment comparisons for safety that are often underpowered. The potential to observe severe safety outcomes may be further limited due to the patients that are enrolled in the trial. To make it straightforward to observe an effect of the prescribed treatments, patients with severe disease, other co-occurring disease, or taking one or more concomitant medications are often excluded from study participation. Further, while the disease under investigation may suggest safety issues likely to occur over the duration of the trial, unplanned safety issues can and often do materialize. This makes it challenging to pre-specify appropriate analyses in advance, with the additional burden of further limiting the available type I error for anticipated events.

Death and disease progression, while important indicators of patient safety, are often analyzed as primary endpoints in clinical trials. Because the strong control of type I error is well understood in these situations, even in the presence of one or more interim analyses, we avoid further discussion specific to these endpoints within this chapter. Here, we focus on the efficient reporting of the considerable volume of safety endpoints that are collected within a clinical trial, with a primary focus on AEs. Because of the limitations described above, the traditional means of data summary—tables and listings—are often ineffective for communicating the story hidden within the data. Data visualization is the key to efficient communication of safety outcomes; we reinforce this idea through the examples below.

Our rationale for the focus on AEs is due to the fact that occurrences of clinically-relevant worsening in other safety endpoints are reported as AEs. For example, significant changes in the laboratory test alanine aminotransferase, an important indicator of liver health, can be represented by the preferred terms Alanine Aminotransferase Abnormal, Alanine Aminotransferase Increased, or Alanine Aminotransferase Decreased when using the Medical Dictionary for Regulatory Activities (MedDRA) (Brown et al. 1999). We illustrate the various methodologies using a clinical trial of patients who experienced an aneurysmal subarachnoid hemorrhage, which is described in Sect. 7.2. Section 7.3 summarizes analysis approaches for safety which are then applied to the sample data in Sect. 7.4. Section 7.5 provides a brief conclusion. All analyses were performed using JMP Clinical 6.1.

7.2 Sample Data

Nicardipine hydrochloride, available in oral and intravenous forms, belongs to the class of calcium channel blockers which are used to treat high blood pressure and angina. Nicardipine was examined in a clinical study of patients experiencing an aneurysmal subarachnoid hemorrhage, which is bleeding between the brain and the tissues that surround the brain (Hayley et al. 1993). The primary endpoint was improvement in patient recovery according to the Glasgow Outcome Scale, with the incidence of cerebral vasospasm, and the incidence of death or disability due to vasospasm serving as important secondary endpoints (Jennett and Bond 1975). The study was a two-week trial in 906 patients randomly assigned to intravenous nicardipine or placebo; 902 patients ultimately received treatment.

The 902 treated patients experienced a total of 4472 treatment emergent AEs (TEAEs), events that occurred on or after the first dose of study drug. Coding with the MedDRA dictionary led to 188 distinct preferred terms contained within 22 system organ classes. This classification of AEs into preferred terms and system organ classes will be used throughout the analyses and figures in this chapter. While the analysis of adverse events hinges on the quality of this coding step, the mechanics and issues surrounding this activity are outside the scope of this chapter.

Please note that the analyses and results summarized here are for illustrative purposes only; no formal conclusions on the safety or effectiveness of nicardipine should be made as a result of this chapter.

7.3 General Considerations for Safety Analyses

7.3.1 Initial Steps

Guideline E2A from the International Conference on Harmonisation (ICH) defines an AE as "any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have to have a causal relationship with this treatment" (ICH 1994). Additionally, serious adverse events (SAEs) are AEs that "result in death, are life threatening, require inpatient hospitalization or prolongation of hospitalization, result in disability or permanent damage, or are congenital anomalies or birth defects" (ICH 1994). Guidelines from the EMA define the additional term adverse drug reaction (ADR) for those events that are viewed by the investigator to have a causal relationship with treatment (EMA 2016).

Adverse events that occur since the previous study visit are reported to the clinician by the patient or care-giver. Additional AEs may be identified by the clinician through in-clinic or laboratory assessments that have worsened since baseline. Details on the severity or toxicity grade (perhaps using the National Cancer Institute's Common Terminology Criteria of Adverse Events, NCI-CTCAE), seriousness, outcome,

duration, the action taken with study drug due to the event, and the investigator's opinion on the relationship to study medication are recorded (NCI 2010). Verbatim event text is coded using MedDRA to maintain consistency in the reporting and grouping of AEs within and across studies and development programs. AEs are traditionally summarized by preferred terms, and grouped by system organ class in order of decreasing frequency of occurrence. In MedDRA Version 19.1, released in September 2016, 22,210 preferred terms are grouped within 27 system organ classes. The binary outcomes as to whether a patient experienced a particular AE or not, are often reported using a risk difference ($\hat{p}_{tj} - \hat{p}_{cj}$), risk ratio ($\hat{p}_{tj}/\hat{p}_{cj}$), or odds ratio $\left(\hat{p}_{tj}\left(1-\hat{p}_{cj}\right)/\left(1-\hat{p}_{tj}\right)\hat{p}_{cj}\right)$, where $\hat{p}_{ij}$ is the probability of experiencing event j of J possible AEs for treatment i (Chuang-Stein et al. 2014; Zhou et al. 2015). Pros and cons for the various measures are discussed in Zhou et al. (2015). This chapter presents risk differences throughout.

Given the large number of potential comparisons of treatment arms for adverse events, Crowe and co-authors suggested a 3-tier approach for the analysis of AEs (Crowe et al. 2009). Pre-planned hypotheses for Tier I events, those AEs expected to occur or of considerable clinical relevance for the disease, would typically not receive adjustment for multiple comparisons unless there were numerous Tier 1 events to consider. Treatment comparisons for unexpected but commonly-occurring (4 or more patients in a single treatment arm) Tier 2 events should consider multiple comparisons. Tier 3 events (those not in Tiers 1 or 2) are rare and should be summarized in a listing. Appropriate multiplicity adjustment for Tier 1 (if required) and Tier 2 events should achieve a reasonable balance between committing type I errors without overly sacrificing the power to detect potential safety signals. The False Discovery Rate (FDR) provides a more balanced approach between type I error and power, since it does not control the familywise error rate (Benjamini and Hochberg 1995). The FDR, typically pre-specified at $\alpha = 0.05$, is the proportion of erroneous rejections among the rejected null hypotheses from a set of multiple tests. In general, with J treatment comparisons of ordered (smallest to largest) p-values $p_{(j)}$, the FDR p-value for the jth hypothesis is

$$p^*_{(j)} = \begin{cases} p_{(J)} & \text{for } j = J \\ \min\left(p^*_{(j)}, \frac{j}{(j-1)} p_{(j-1)}\right) & \text{for } j = 1, 2, \ldots (J-1) \end{cases}$$

Corresponding simultaneous 95% FDR confidence intervals can be defined by finding the largest j where $p_{(j)} \leq j\alpha/J$ and using $\alpha^* = j\alpha/J$ for all J confidence intervals (2005). An alternate FDR methodology, the double FDR, could also be considered to account for the relationship among AEs through a grouping variable such as system organ class (Mehrotra and Adewale 2012).

7.3.2 Further Analyses

Given the numerous characteristics present for AEs, it is worthwhile to perform additional analyses to determine the impact of event seriousness, severity, or causality on safety assessments between the study treatments. For example, in trials of oncology, it is recommended to summarize AE incidence for all grades as well as for those events that are considered more severe, NCI-CTCAE grade 3 and above (EMA 2016). Similar analyses are often presented for the subset of events determined to be ADRs. Additionally, summaries of AEs may be performed for different stages of the trial, describing safety before, during, and after treatment.

However, apart from summarizing events for different trial stages, the influence of time tends to be ignored in most presentations of AEs. Since patients with longer follow-up have greater opportunity to experience one or more safety outcomes, it is important to consider exposure-adjusted incidence rates or time-to-first events, particularly in studies with varying patient exposure (Koch et al. 1993; Liu et al. 2006; Stokes et al. 2012; Zhou et al. 2015; Allignol et al. 2016; EMA 2016; Proctor and Schumacher 2016). However, these methodologies do have limitations. Time-to-first event analyses are limited in that they only describe when the first event occurs. Exposure-adjusted incidence rates assume a constant hazard rate across time. Liu and co-authors (2006) suggest that this assumption is likely to hold for rare events, though it should be assessed in practice since this expectation may not apply for many events. Breaking the study period up into meaningful mutually-exclusive time intervals allows for the possibility of constant hazards to hold within smaller time intervals.

In general, however, analyses within time intervals can provide a more informative analysis that makes it possible to view how the risk of AEs changes over the course of a clinical trial. For example, the risk of certain events may reduce as patients develop tolerability to the study medications. Alternatively, greater exposure to drug may result in an increased likelihood of certain events. Zink et al. (2013) illustrates how multiple plots or animation can be used to communicate the instantaneous risk within time intervals. Similar presentations can be used to present analyses of cumulative risk over time. For example, guidance suggests presentations of cumulative AE rates for oncology studies at 3, 6, and 12 months, with the addition of other time points depending on the underlying nature of the disease and the duration of the trial (EMA 2016). Presenting AEs by time intervals serves an additional purpose, since differential rates of drop out between the treatment arms can spawn misleading results for the entire treatment period. After all, patients responding to treatment with longer follow-up times have greater opportunity to experience one or more safety outcomes. Presentations of instantaneous risk by time interval is also one way to account for and summarize the recurrence of events observed during the clinical trial, though more formal analyses to assess the average number of events experienced over time are available (Johnston and So 2003; Nelson 2003; Diao et al. 2015; Hengelbrock et al. 2016). Finally, Koch et al. (1993) present a large-sample method to summarize

the total number of events experienced accounting for the correlation between event frequency and patient exposure.

Given the rarity of many individual events, an alternative strategy to identify safety signals is to analyze groupings of preferred terms by analyzing higher level terms or higher level group terms from the MedDRA hierarchy. In MedDRA 19.1, there are 1732 and 335 higher level and higher level group terms, respectively. Similarly, the analyst could examine the incidence of standardised MedDRA queries (SMQs), which are groups of lower level and preferred terms that describe a particular medical condition (Mozzicato 2007). MedDRA 19.1 documents 217 SMQs. Analyzing higher-level MedDRA terms or other groups of events may not allow us to make conclusions about individual events at the preferred term level, but it may signal the potential for increased risk of contributing AEs once sufficient sample size is obtained. Other safety analyses consider the co-occurrence of sets of events observed on study, without the formal classifications of a medical dictionary (DuMouchel and Pregibon 2001; Goldberg-Alberts and Page 2006).

In lieu of grouping sets of events, it may be worthwhile to examine safety outcomes within more homogeneous groups of patients. Subgroups are frequently considered for the analysis of safety and efficacy endpoints, with 70% of clinical trials reporting at least some results within subgroups (Pocock et al. 2002). Subgroup analyses are beneficial in that they provide clinicians with information on the potential for differential treatment response within important demographic, genetic, disease, environmental, behavioral or regional characteristics (Chuang-Stein et al. 2014; Quan et al. 2010). In addition, recent data-driven methodologies can be used to identify subgroups using combinations of individual factors to characterize sets of patients with differential response to treatment, though the importance of these subgroups need to be confirmed in further study (Battioui et al. 2013; Dusseldorp and Mechelen 2014; Foster et al. 2011; Loh 2011; Lipkovich et al. 2011; Lipkovich and Dmitrienko 2014; Negassa et al. 2005; Su et al. 2009; Zink et al. 2015). For safety outcomes, the exploration of effects within subgroups can identify groups of patients for whom the new therapy may be inappropriate.

From a regulatory perspective, subgroup analyses are important to show that the estimated overall effect is broadly applicable to patients and to assess risk-benefit across the proposed indication, particularly when the study population is heterogeneous (CHMP 2014). Further, examining results within subgroups allows the study team to assess the consistency and robustness of results obtained for the entire study population, as well as to generate hypotheses for future research (Cui et al. 2002). Subgroup analyses would likely be considered for important Tier I events.

When reporting results within subgroups, transparency is key for appropriate interpretation of results. Details on the number of subgroups assessed (not just reported), whether subgroups were determined pre or post hoc, multiplicity adjustments were applied, stratified randomization was used, or heterogeneity was assessed and through what method should be clearly described (Lagakos 2006; Wang et al. 2007). For multiplicity, details as to whether adjusted or unadjusted p-values are presented or simultaneous or unadjusted confidence or credible intervals should be clearly described. However, regulatory guidance appears to prefer presenting unadjusted p-values and

intervals for subgroup analyses as they are "investigations [that] serve as an indicator for further exploration" (CHMP 2014). Even though power tends to be low for tests of interaction, many authors suggest that heterogeneity of treatment effects should always be evaluated, and regulatory guidance encourages reporting estimates and confidence intervals for these interaction tests (Pocock et al. 2002; Lagakos 2006; CHMP 2014). Further, the literature highlights that the presence and the size of interaction depends on the choice of the measure of divergence between the treatment groups (Chuang-Stein et al. 2014; CHMP 2014).

7.4 Safety Analysis of Sample Data

7.4.1 Initial Steps

In this section, we illustrate the effectiveness of data visualization for communicating the results of AE analyses for the nicardipine trial. For example, Fig. 7.1 presents a volcano plot to summarize the incidence of adverse events (Zink et al. 2013). The x-axis represents the nicardipine minus placebo risk difference while the y-axis represents the –log10 transformation of the unadjusted p-value from a Cochran-Mantel-Haenszel correlation statistic. The smaller the p-value, the larger the value on the y-axis; y can be thought of as the number of decimal places or number of zeros in the p-value derived from the comparison of risk between the treatments. References lines are drawn to show significant events with no-adjustment ($-\log10(0.05) = 1.3$) or FDR adjustment ($-\log10(0.0013) = 2.876$, where $\alpha^* = 5/188 \times 0.05 = 0.0013$); events are significant if the center of the bubble is above a particular reference line. Given the location of events relative to the 0 point (no difference) on the horizontal axis, Vasoconstriction and Hypertension exhibit elevated risk on placebo, while Phlebitis, Hypotension, and Isosthenuria have elevated risk on nicardipine. Though an asymptotic test is used here for illustration purposes, analysts should consider using exact methods when sample sizes are small or when events are rare, and should consider a test that accounts for the stratification applied to treatment randomization. Bubble area is proportional to the total number of patients on either treatment that experience a particular adverse event. Of the five signals identified in Fig. 7.1, Vasoconstriction and Isosthenuria were experienced by the most and least patients, respectively. Bubble color communicates the system organ class, red for the Vascular Disorders of Hypotension, Hypertension, Phlebitis, and Vasoconstriction, and green for the Renal and Urinary Disorder of Isosthenuria. An alternate approach could color bubbles according to event tier. Figure 7.2 is identical to Fig. 7.1 except that the bubble size is proportional to the inverse of the variance of the treatment difference $\left(\frac{p_n(1-p_n)}{n_n} + \frac{p_p(1-p_p)}{n_p}\right)^{-\frac{1}{2}}$. This plot helps the team assess the amount of information available relative to the magnitude of the treatment effect for safety signals.

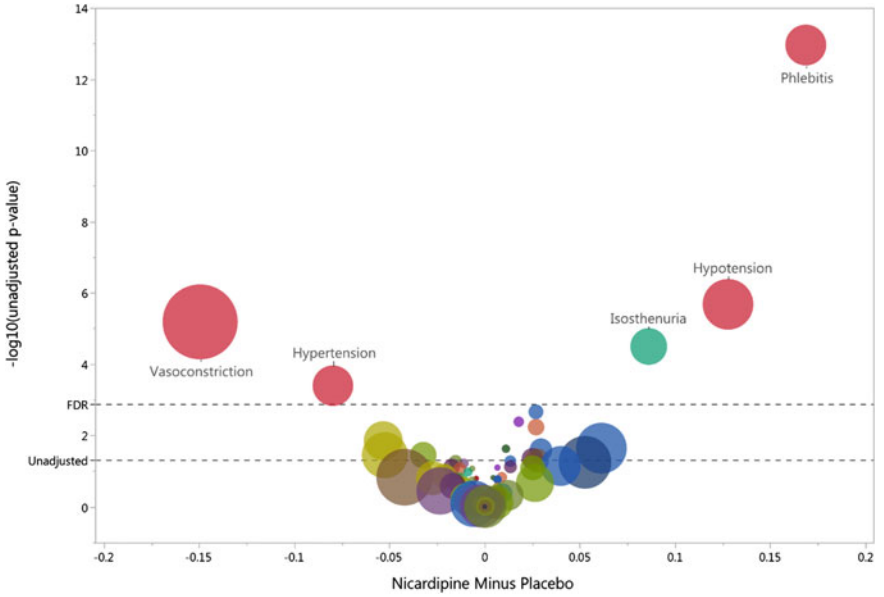


Fig. 7.1 Volcano plot comparing the proportion of treatment emergent adverse events between treatments with bubbles sized according to event frequency. Unadjusted reference line drawn at $-\log10(0.05) = 1.3$. FDR reference line drawn at $-\log10(0.0013) = 2.876$, where $\alpha^* = 5/188 \times 0.05 = 0.0013$. Alternatively, the FDR reference line could be drawn at $-\log10$(maximum unadjusted p-value $\leq \alpha^*$) as in Zink et al. (2013). Bubble areas are proportional to the total number of patients that experience an adverse event for both treatments combined. Bubbles are colored according to system organ class with red or green bubbles referring to vascular disorders or renal and urinary disorders, respectively

Now that we have identified differential safety responses between the treatments in Figs. 7.1 and 7.2, we can explore these events in greater detail. For example, Fig. 7.3 presents a forest plot and dot plot to communicate the variability around treatment effects and the individual event incidence (Amit et al. 2008). Here, blue intervals highlight the events with greater risk for nicardipine, while red highlights events with greater risk for placebo. Further, forest plots help communicate the sensitivity of findings within important demographic subgroups; Fig. 7.4 presents an analysis of Isosthenuria within subgroups. The left panel summarizes unadjusted 95% confidence intervals which highlights elevated risk for Isosthenuria across most demographic subgroups. Based on recommendations from the CHMP (2014), interaction tests are summarized using a forest plot in the right panel and are based on unadjusted 95% confidence intervals for the difference in treatment effects between the two subgroup levels (level 1 minus level 2). Finally, the heat map in Fig. 7.5 assesses the sensitivity of analysis findings by examining the standardized effect $(p_n - p_c)\left(\frac{p_n(1-p_n)}{n_n} + \frac{p_p(1-p_p)}{n_p}\right)^{-\frac{1}{2}}$ across SAEs, events of varying levels of severity, and events with varying levels of causality to drug. Darker blue highlights the sce-

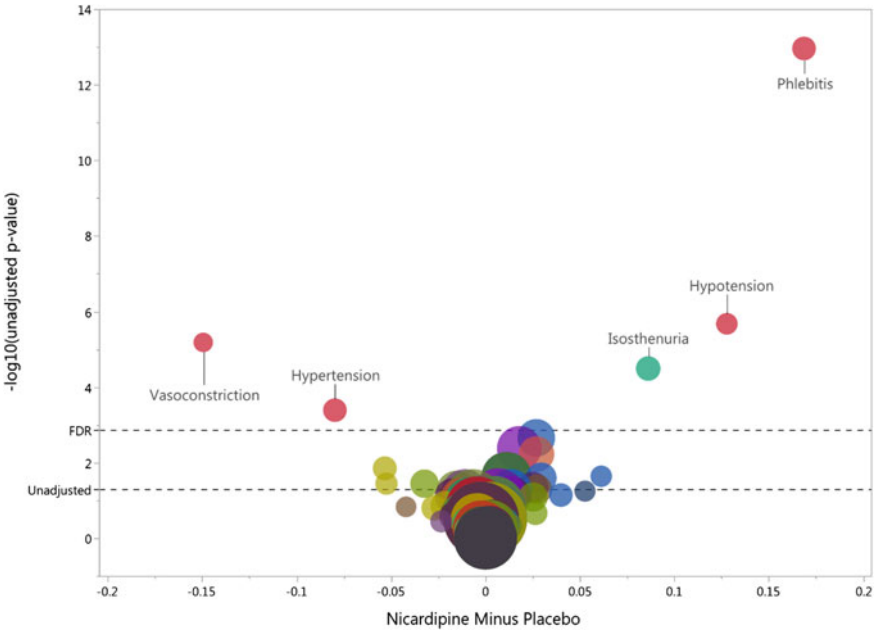


Fig. 7.2 Volcano Plot Comparing the Proportion of Treatment Emergent Adverse Events Between Treatments with Bubbles Sized According to Variability. Unadjusted reference line drawn at $-\log10(0.05) = 1.3$. FDR reference line drawn at $-\log10(0.0013) = 2.876$, where $\alpha^* = 5/188 \times 0.05 = 0.0013$. Bubble areas are proportional to the inverse of the variance of the treatment difference $\left(\frac{p_n(1-p_n)}{n_n} + \frac{p_p(1-p_p)}{n_p}\right)^{-\frac{1}{2}}$. Bubbles are colored according to system organ class with red or green bubbles referring to vascular disorders or renal and urinary disorders, respectively

narios with greater risk for nicardipine, while darker red highlights scenarios with greater risk for placebo. However, it is important to not over-interpret the results in Fig. 7.5, since reduced risk (colors closer to gray) may be due to scenarios with a greatly reduced number of events.

7.4.2 Accounting for Time and Patient Exposure

Up until now, there has been no consideration for time or exposure in the analysis of AEs. We being our assessment of the effects of time on safety with Fig. 7.6, which summarizes the instantaneous risk of events with 3 day intervals (Zink et al. 2013). To improve the presentation, only events that are significant in at least one interval using the FDR reference are displayed. Over the 4 intervals, one can observe a steady reduction or increase in the risk of Hypotension or Phlebitis, respectively, when comparing nicardipine to placebo. Further, there is one period (days 7–9) where there

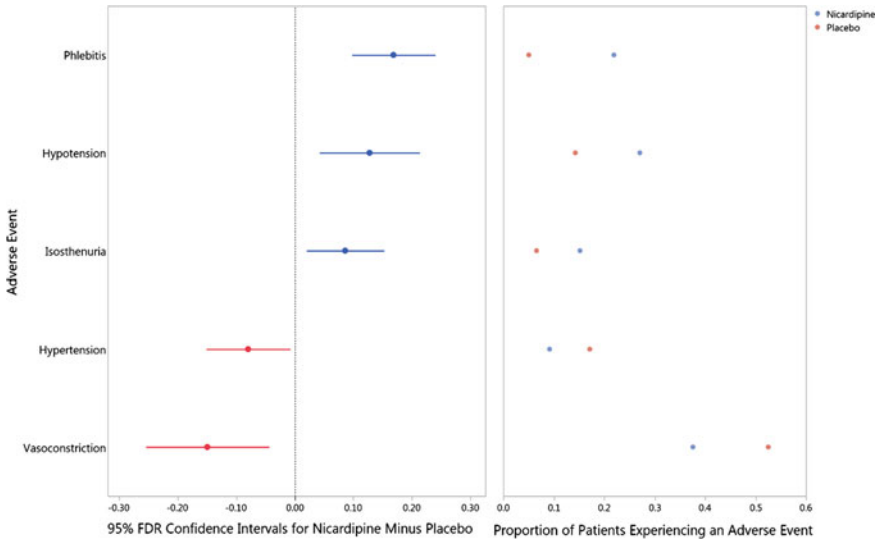


Fig. 7.3 FDR confidence intervals and event incidence for identified safety signals. Presentation suggested as in Amit et al. (2008). Left panel displays a forest plot of FDR intervals for nicardipine minus placebo for signals identified from Fig. 7.1. Reference line is drawn at 0 to indicate no difference between nicardipine and placebo, with blue or red intervals showing elevated risk on nicardipine or placebo, respectively. Right panel presents a dot plot to communicate the incidence of each AE for each treatment arm

appears to be some elevation of Intracranial Pressure Increased in placebo compared to nicardipine by FDR; this event was not observed in Figs. 7.1 and 7.2. However, if Intracranial Pressure Increased had been a Tier I event, there would appear to be some evidence of increased risk for placebo across all time intervals. Though informative on its own merits, Fig. 7.6 may suggest further scrutiny of some AEs with additional exploratory analyses. It is important to note that Fig. 7.6 clearly demonstrates the importance for checking the constant hazards assumption for exposure-adjusted incidence rates. As an aside, a similar presentation could summarize the cumulative risk of events across time.

As an alternative approach to accounting for time, Fig. 7.7 summarizes a Kaplan-Meier analysis of time-to-first event, summarizing the log-rank or Wilcoxon test along the y-axis in the left or right panel, respectively. Here, the x-axis represents the maximum distance (nicardipine minus placebo) between the two Kaplan-Meier curves, though other measures can be used to emphasize the differences between the treatments. Both analyses identify the same five signals that were identified above, with the inclusion of Pleural Effusion for the log-rank test. Which test should be used in practice? In some instances, it may not be clear which test may be most appropriate for a given data set. The log-rank test assumes proportional hazards between the two treatments which may not hold; the Wilcoxon test places greater emphasis on earlier event times (Collett 2015). Ultimately, whichever test is selected

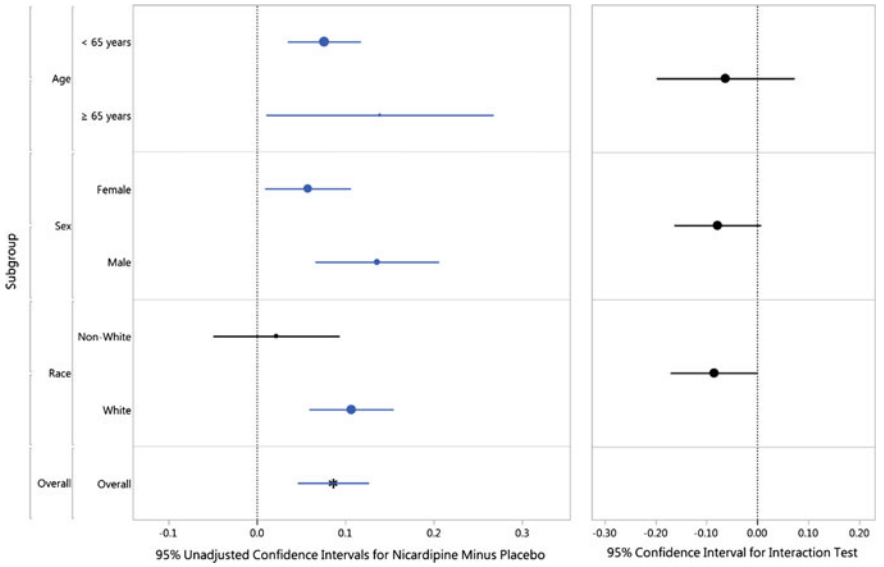


Fig. 7.4 Subgroup analysis of Isosthenuria. Unadjusted 95% confidence intervals are based on the risk difference of isosthenuria for nicardipine minus placebo using a normal approximation, with blue intervals highlighting events with elevated risk for nicardipine. Interaction tests are based on unadjusted 95% confidence intervals for the difference in treatment effects between the two subgroup levels (level 1 minus level 2). Bubble areas for subgroups in the left panel are proportional to the total number of patients within each subgroup level

for the primary analysis should be pre-specified, with the other test serving as an important sensitivity analysis. Alternatively, a bivariate test of log-rank and Wilcoxon scores can be applied (Tangen and Koch 1999). Finally, event recurrence can be assessed using similar plots. For example, a proportional means model can be used to compare the mean cumulative function between the treatments (Johnston and So 2003).

7.4.3 Standardised MedDRA Queries

We briefly mentioned SMQs above. Recall, that SMQs are groups of lower level and preferred terms that describe a particular medical condition. For example, the preferred terms Agitation, Delirium, Disorientation, Hallucination, and Psychotic Disorder contribute to the SMQ Dementia. Here, our goal is to potentially gain power for statistical comparisons by combining related events in order to describe a particular disease state or syndrome. It may not be possible to make formal conclusions about the SMQ itself (say, to report in a drug label), but these analyses provide insight into the ways in which the treatments affect various aspects of safety. Potentially,

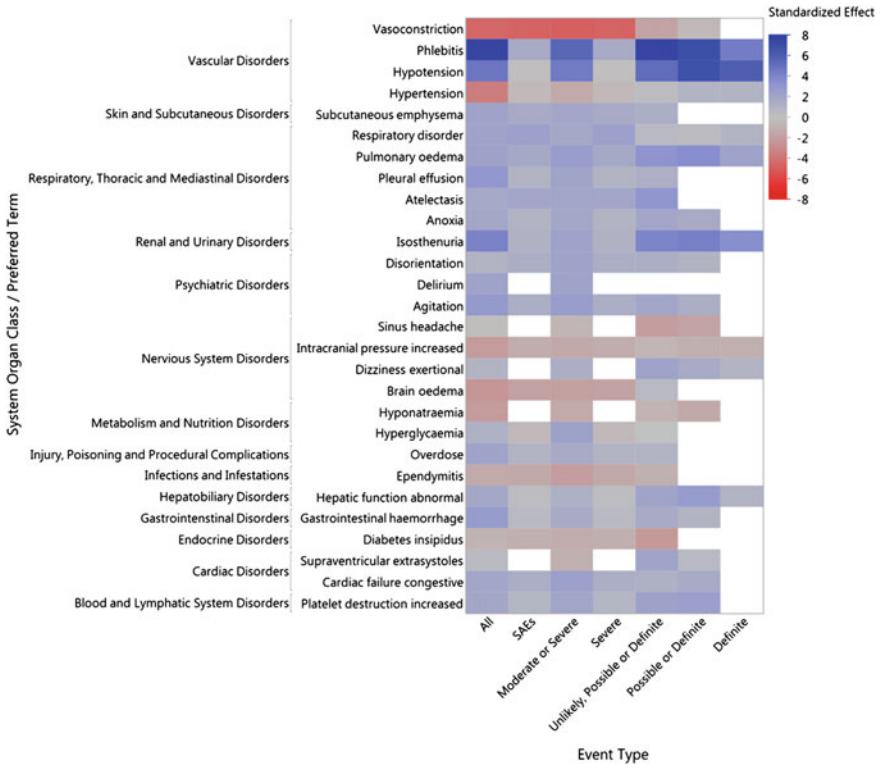


Fig. 7.5 Sensitivity analysis comparing the proportion of treatment emergent adverse events between treatments. The standardized effect is $(p_n - p_c)\left(\frac{p_n(1-p_n)}{n_n} + \frac{p_p(1-p_p)}{n_p}\right)^{-\frac{1}{2}}$. Darker blue or red indicates higher risk on nicardipine or placebo, respectively. Cells are white when the standardized effect cannot be calculated, most often when no events occur. Due to space limitations, only the events with at least one significant unadjusted p-value for any analysis are presented

differential SMQ response between study treatments may serve as an early warning for individual contributing events once sufficient data are accumulated. Searches for SMQs can take various forms: narrow, which limits the set of terms to those most likely to identify patients with a given condition, or broad, which contains additional terms in order to "cast a wide net". Further, in MedDRA Version 19.1, there are 10 SMQs with algorithms, which often amounts to accumulating a sufficient number and variety of events of various subtypes. There are 25 hierarchies among the SMQs, though not all queries have a hierarchical relationship with other queries. Here, we pay no consideration to the hierarchical relationships of observed SMQs. Figure 7.8 summarizes the frequency of 58 SMQs that were observed based on an analysis of treatment emergent preferred terms.

Similar to Fig. 7.1, Fig. 7.9 contains a volcano plot summarizing the difference in the incidence of SMQs between nicardipine and placebo. Here, bubbles are colored

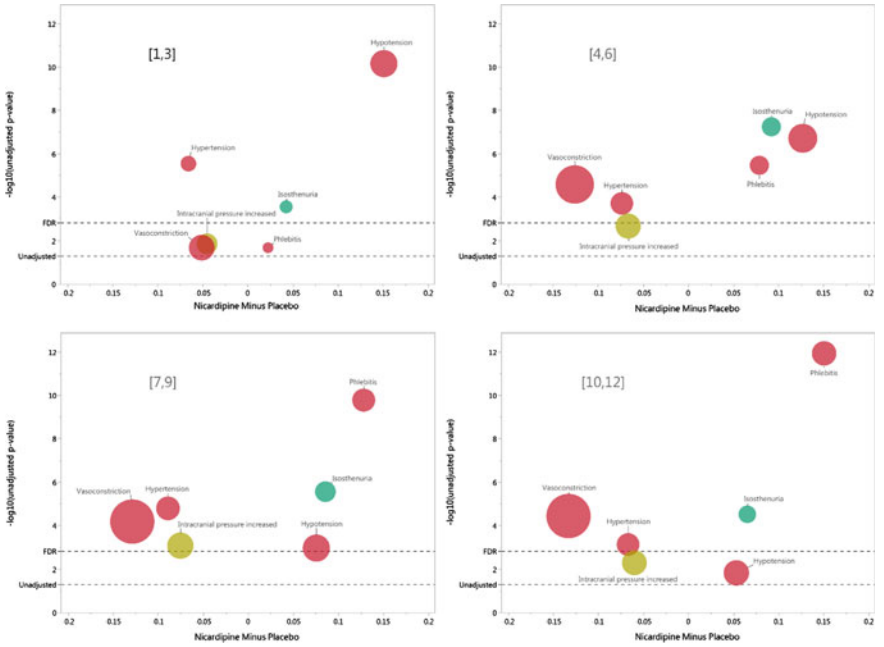


Fig. 7.6 Volcano Plot Comparing the Proportion of Treatment Emergent Adverse Events Between Treatments Across Time Intervals. Unadjusted reference line drawn at $-\log10(0.05) = 1.3$. FDR reference line drawn at $-\log10(0.0015) = 2.8297$, where $\alpha^* = 18/608 \times 0.05 = 0.0015$. Bubble areas are proportional to the total number of patients that experience an adverse event for both treatments combined. Bubbles are colored according to system organ class. Starting at the upper-left hand corner, volcano plots summarize the incidence of events in time intervals for study days 1–3, 4–6, 7–9 and 10–12. Only events that are significant in at least one interval using the FDR reference are displayed

according to the number of distinct preferred terms that contribute to each query. As mentioned above, Dementia has 5 preferred terms associated with it, which is why it is a light blue color. For the other query signals: Thrombophlebitis is composed of preferred terms Thrombophlebitis and Phlebitis; Hostility/aggression is composed of Agitation, Paranoia, Personality Disorder, and Psychotic Disorder; and Anaphylactic Reaction is composed of Cardiac Arrest, Choking, Cough, Dyspnoea, Hyperventilation, Hypotension, and Laryngeal Oedema. This raises an interesting question, is it possible to apply the 3-tier system described in Sect. 7.3.1 to SMQs? One recommendation is that any SMQs that contains a Tier 1 event would itself be considered Tier 1, with all other SMQs of sufficient numbers relegated to Tier 2. However, given that these analyses are exploratory beyond the testing of individual event terms, applying an FDR correction across all queries may be appropriate. However, as we have observed in this simple example, the terms Agitation and Psychotic Disorder contribute to both Dementia and Hostility/aggression, which creates a dependency among the individual tests. In these cases, FDR methods that formally consider the

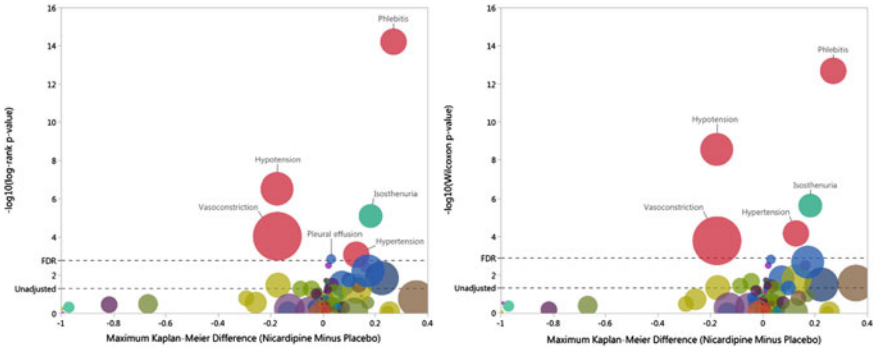


Fig. 7.7 Volcano plot comparing the time-to-first-event for treatment emergent adverse events between treatments with bubbles sized according to event frequency. Kaplan-Meier analysis of time-to-first event summarizing the log-rank or Wilcoxon tests on the y-axis in the left or right panel, respectively. The x-axis represents the maximum distance (nicardipine minus placebo) between the Kaplan-Meier curves. Unadjusted reference line drawn at $-\log10(0.05) = 1.3$. FDR reference line for the log-rank test is drawn at $-\log10(0.0016) = 2.7970$, where $\alpha^* = 6/188 \times 0.05 = 0.0016$. FDR reference line for the Wilcoxon test is drawn at $-\log10(0.0013) = 2.8762$, where $\alpha^* = 5/188 \times 0.05 = 0.0013$. Bubble areas are proportional to the total number of patients that experience an adverse event for both treatments combined. Bubbles are colored according to system organ class

association of multiple tests may provide a more appropriate adjustment for multiple comparisons (Yekutieli and Benjamini 1999; Benjamini and Yekutieli 2001). Finally, a similar presentation in Fig. 7.5 could be applied to SMQs, subsetting to specific events that meet various criteria. Further, additional columns could be added to summarize the findings across broad or algorithmic queries.

7.5 Conclusions

This chapter summarized analysis and reporting approaches for adverse events, and illustrated methodologies using data from a sample clinical trial. Due to space considerations, several issues were not addressed in the text above. These points are briefly listed here so that interested readers can explore these topics in greater detail in the references that follow.

- It is important to proactively plan for a comprehensive safety evaluation at the start of any development program, a plan that considers the underlying challenges of the disease, as well as the unique features of treatment and patient management. Though not a regulatory requirement, the Safety Planning, Evaluation and Reporting Team (SPERT) recommends a Program-wide Safety Analysis Plan (PSAP) to document the statistical aspects of safety during clinical development and post-marketing activities (Crowe et al. 2009).

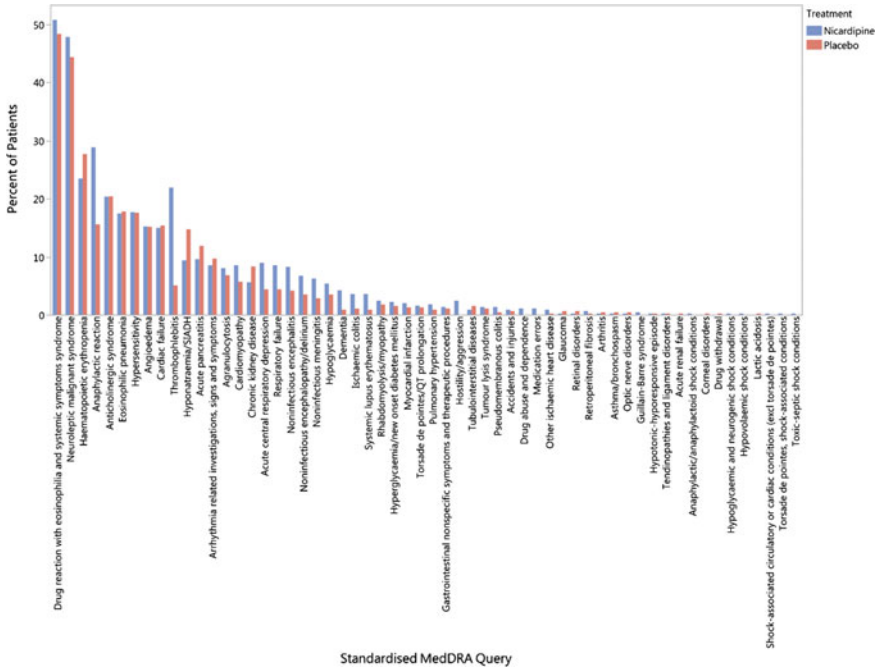


Fig. 7.8 Bar chart summarizing the percent of patients with treatment emergent standardised MedDRA queries by treatment. Standardised MedDRA queries are defined using only preferred terms using a narrow search at the individual query level

- The methods described attempt to identify population shifts in important safety parameters between the treatments. Ethically, it is necessary to identify individual patients experiencing severe outcomes in order for them to receive appropriate care. The EudraVigilance Expert Working Group maintains a list of important medical events (IMEs) for which it may be important to screen AE data for the presence of individual cases (EudraVigilance 2016a Aug, b Sept). Ongoing screening of serious and unexpected suspected adverse reactions is suggested in the Federal Register and the recent draft FDA guidance on safety assessment (US FDA 2010, 2015).
- All therapies carry some level of risk, and for more severe diseases, patients may be more willing to accept a greater degree of toxicity in order to obtain an important benefit than they would be for less grievous conditions. Balancing the potential benefits and risks of new therapies is challenging, and is an area of active research (Jiang and He 2016; Bender et al. 2016).
- The rarity of many safety endpoints will require a meta-analysis of multiple studies for sufficient power to generate meaningful inference for the safety population, as well as more precise estimates of the treatment response within various subgroups. (Koch et al. 1993; Crowe et al. 2009; Berlin et al. 2012; Chuang-Stein et al. 2014).

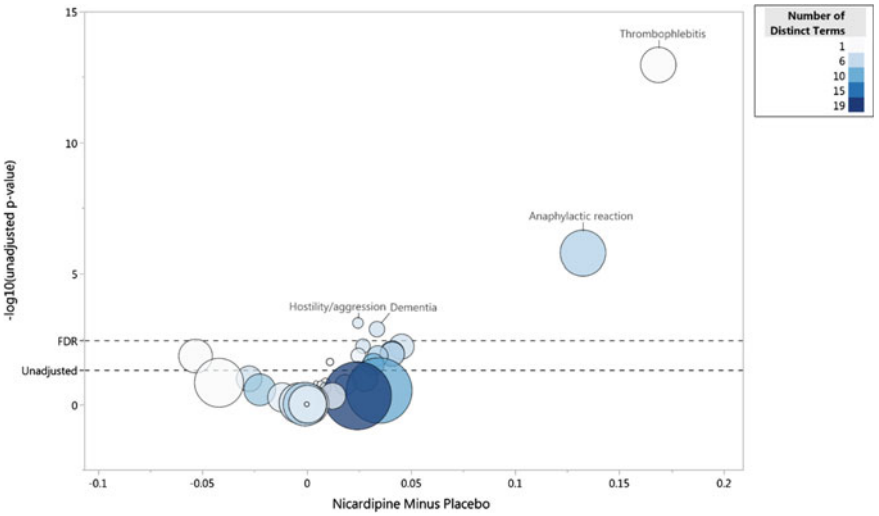


Fig. 7.9 Volcano plot comparing the proportion of treatment emergent standardised MedDRA queries between treatments with bubbles sized according to query frequency. Unadjusted reference line drawn at $-\log10(0.05) = 1.3$. FDR reference line drawn at $-\log10(0.0034) = 2.4624$, where $\alpha^* = 4/58 \times 0.05 = 0.0034$. Bubble areas are proportional to the total number of patients that experience a standardised MedDRA query for both treatments combined. Bubbles are colored according to the number of distinct preferred terms that are observed for each query. Here, standardised MedDRA queries are defined using only preferred terms and using a narrow search at the individual query level

Meta-analyses should be pre-planned in the PSAP and assess the heterogeneity and poolability of the included clinical trials, not simply reflect a naïve grouping of patients from multiple studies, ignoring the variability in treatment effects between studies.

Though AEs are an important part of the safety assessment for any new drug, device, or biologic, there are numerous other safety endpoints to consider. Readers interested in greater detail on the analysis and reporting of AEs and other safety outcomes can explore texts by Jiang and Xia (2014) or Gould (2015), or revisit Gilbert (1993). For greater therapeutic focus, readers can review a recent examination of safety specific to clinical trials in oncology (Ivanova et al. 2017). See Turner et al. (2017) for an in-depth overview of cardiovascular safety. Finally, this chapter emphasized the importance of data visualization for summarizing, interpreting and communicating analyses of safety outcomes. Those individuals interested in additional graphical presentations of safety data can review Chuang-Stein et al. (2001), Amit et al. (2008), Krause and O'Connell (2012), Duke et al. (2015), or Matange (2016).

Acknowledgements The author thanks Karl Peace for the invitation to contribute to this volume.

References

Allignol, A., Beyersmann, J., & Schmoor, C. (2016). Statistical issues in the analysis of adverse events in time-to-event data. *Pharmaceutical Statistics, 15,* 297–305.

Amit, O., Heiberger, R. M., & Lane, P. W. (2008). Graphical approaches to the analysis of safety data from clinical trials. *Pharmaceutical Statistics, 7,* 20–35.

Battioui, C., Shen, L., & Ruberg, S. J. (2013). A resampling-based ensemble tree method to identify patient subgroups with enhanced treatment effect. In *Proceedings to the Joint Statistical Meetings*.

Bender, R., Beckmann, L., & Lange, S. (2016). Biometrical issues in the analysis of adverse events within the benefit assessment of drugs. *Pharmaceutical Statistics, 15,* 292–296.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57,* 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29,* 1165–1188.

Benjamini, Y., & Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association, 100,* 71–81.

Berlin, J. A., Crowe, B. J., Whalen, E., Xia, H. A., Koro, C. E., & Kuebler, J. (2012). Meta-analysis of clinical trial safety data in a drug development program: answers to frequently asked questions. *Clinical Trials, 10,* 20–31.

Brown, E. G., Wood, L., & Wood, S. (1999). The medical dictionary for regulatory activities (MedDRA). *Drug Safety, 20,* 109–117.

Chuang-Stein, C., Le, V., & Chen, W. (2001). Recent advancements in the analysis and presentation of safety data. *Drug Information Journal, 35,* 377–397.

Chuang-Stein, C., Li, Y., Kawai, N., Komiyama, O., & Kuribayashi, K. (2014). Detecting safety signals in subgroups. In Q. Jiang & H. A. Xia (Eds.), *Quantitative evaluation of safety in drug development: Design, analysis and reporting*. Boca Raton, Florida: CRC Press.

Collett, D. (2015). *Modelling survival data in medical research* (3rd ed.). Boca Raton, Florida: CRC Press.

Committee for Medicinal Products for Human Use (CHMP). (2014). Guideline on the investigation of subgroups in confirmatory clinical trials (draft). European Medicines Agency. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf.

Crowe, B. J., Xia, H. A., Berlin, J. A., Watson, D. J., Shi, H., Lin, S. L., et al. (2009). Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: A report of the safety planning, evaluation, and reporting team. *Clinical Trials, 6,* 430–440.

Cui, L., Jung, H. M. J., Wang, S. J., & Tsong, Y. (2002). Issues related to subgroup analysis in clinical trials. *Journal of Biopharmaceutical Statistics, 12,* 347–358.

Diao, L., Cook, R. J., & Lee, K. A. (2015). Statistical analysis of recurrent adverse events. In A. L. Gould (Ed.), *Statistical methods for evaluating safety in medical product development*. Wiley Ltd.: Chichester, United Kingdom.

Duke, S. P., Bancken, F., Crowe, B., Soukup, M., Botsis, T., & Forshee, R. (2015). Seeing is believing: good graphic design principles for medical research. *Statistics in Medicine, 34,* 3040–3059.

DuMouchel, W., & Pregibon, D. (2001). Empirical Bayes screening for multi-item associations. In *Proceedings of Knowledge Discovery and Data Mining International Conference* (pp. 67–76).

Dusseldorp, E., & Mechelen, I. V. (2014). Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. *Statistics in Medicine, 33,* 219–237.

EudraVigilance Expert Working Group. (2016a, September). Important Medical Event Terms List (MedDRA version 19.1). http://www.ema.europa.eu/ema/pages/includes/document/open_document.jsp?webContentId=WC500208836.

EudraVigilance Expert Working Group. (2016b, August). Inclusion/exclusion criteria for the 'Important Medical Events' list. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2016/08/WC500212100.pdf.

European Medicines Agency. (2016). Draft guideline on the evaluation of anticancer medicinal products in man. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2016/03/WC500203320.pdf.

Foster, J. C., Taylor, J. M. G., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine, 30,* 2867–2880.

Gilbert, G. S. (Ed.). (1993). *Drug safety assessment in clinical trials*. New York: Marcel Dekker.

Goldberg-Alberts, R., & Page, S. (2006). Multivariate analysis of adverse events. *Drug Information Journal, 40,* 99–110.

Gould, A. L. (Ed.). (2015). *Statistical methods for evaluating safety in medical product development*. Chichester, United Kingdom: Wiley Ltd.

Haley, E. C., Kassell, N. F., & Torner, J. C. (1993). A randomized controlled trial of high-dose intravenous nicardipine in aneurysmal subarachnoid hemorrhage. *Journal of Neurosurgery, 78,* 537–547.

Hengelbrock, J., Gillhaus, J., Kloss, S., & Leverkus, F. (2016). Safety data from randomized controlled trials: Applying models for recurrent events. *Pharmaceutical Statistics, 15,* 315–323.

International Conference on Harmonisation. (1994). Guideline E2A: Clinical Safety Data Management: Definitions and Standards for Expedited Reporting. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E2A/Step4/E2A_Guideline.pdf.

Ivanova, A., Marchenko, O., Jiang, Q., & Zink, R. C. (2017). Safety monitoring and analysis in oncology trials. In S. Roychoudhury & S. Lahiri (Eds.), *Statistical challenges in oncology clinical development*. Boca Raton, Florida: CRC Press (Forthcoming).

Jennett, B., & Bond, M. (1975). Assessment of outcome after severe brain damage: A practical scale. *Lancet, 1,* 480–484.

Jiang, Q., & He, W. (Eds.). (2016). *Benefit-risk assessment methods in medical product development*. Boca Raton, Florida: CRC Press.

Jiang, Q., & Xia, H. A. (Eds.). (2014). *Quantitative evaluation of safety in drug development: Design, analysis and reporting*. Boca Raton, Florida: CRC Press.

Johnston, G., & So, Y. (2003). Analysis of data from recurrent events. *SAS User Group International, Statistics and Data Analysis, 28,* 1–12.

Koch, G. G., Schmid, J. E., Begun, J. M., & Maier, W. C. (1993). Meta-analysis of drug safety data. In G. S. Gilbert (Ed.), *Drug safety assessment in clinical trials*. New York: Marcel Dekker.

Krause, A., & O'Connell, M. (Eds.). (2012). *A picture is worth a thousand tables: Graphics in life sciences*. New York: New York Springer.

Lagakos, S. (2006). The challenge of subgroup analyses: Reporting without distorting. *New England Journal of Medicine, 354,* 1667–1669.

Lipkovich, I., & Dmitrienko, A. (2014). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. *Journal of Biopharmaceutical Statistics, 24,* 130–153.

Lipkovich, I., Dmitrienko, A., Denne, J., & Enas, G. (2011). Subgroup identification based on differential effect search—A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine, 30,* 2601–2621.

Liu, G. F., Wang, J., Liu, K., & Snavely, D. B. (2006). Confidence intervals for an exposure adjusted incidence rate difference with applications to clinical trials. *Statistics in Medicine, 25,* 1275–1286.

Loh, W. Y. (2011). Classification and regression trees. *Data Mining and Knowledge Discovery, 1,* 14–23.

Matange, S. (2016). *Clinical graphs using SAS*. Cary, North Carolina: SAS Institute Inc.

Mehrotra, D. V., & Adewale, A. J. (2012). Flagging clinical adverse experiences: Reducing false discoveries without materially compromising power for detecting true signals. *Statistics in Medicine, 31,* 1918–1930.

Mozzicato, P. (2007). Standardised MedDRA queries: Their role in signal detection. *Drug Safety, 30,* 617–619.

National Cancer Institute. (2010). Common Terminology Criteria for Adverse Events (CTCAE) v4.0. http://evs.nci.nih.gov/ftp1/CTCAE/About.html.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., & Boivin, J. F. (2005). Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria. *Statistical Computing, 15,* 231–239.

Nelson, W. (2003). Recurrent-events data analysis for repairs, disease episodes, and other applications. In *ASA-SIAM Series on Statistics and Applied Probability*. Philadelphia, PA: SIAM.

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine, 21,* 2917–2930.

Proctor, T., & Schumacher, M. (2016). Analyzing adverse events by time-to-event models: The CLEOPATRA study. *Pharmaceutical Statistics, 15,* 306–314.

Quan, H., Mingyu, L., Chen, J., Gallo, P., Binkowitz, B., Ibia, E., et al. (2010). Assessment of consistency of treatment effects in multiregional clinical trials. *Drug Information Journal, 44,* 617–632.

Stokes, M. E., Davis, C. S., & Koch, G. G. (2012). *Categorical data analysis using SAS* (3rd ed.). Cary, North Carolina: SAS Institute Inc.

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research, 10,* 141–158.

Tangen, C. M., & Koch, G. G. (1999). Nonparametric analysis of covariance for hypothesis testing with logrank and Wilcoxon scores and survival-rate estimation in a randomized clinical trial. *Journal of Biopharmaceutical Statistics, 9,* 307–338.

Turner, J. R., Karnad, D. R., & Kothari, S. (2017). *Cardiovascular safety in drug development and therapeutic use: New methodologies and evolving regulatory landscapes*. Cham, Switzerland: Springer.

US Food and Drug Administration. (2010, September). Final rule, investigational new drug safety reporting requirements for human drug and biologic products and safety reporting requirements for bioavailability and bioequivalence studies in humans. Federal Register. https://www.gpo.gov/fdsys/pkg/FR-2010-09-29/pdf/2010-24296.pdf.

US Food and Drug Administration. (2015). Guidance for industry: Safety assessment for IND safety reporting (draft).

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., & Drazen, J. M. (2007). Statistics in medicine—Reporting of subgroup analyses in clinical trials. *New England Journal of Medicine, 357,* 2189–2194.

Yekutieli, D., & Benjamini, Y. (1999). Resampling based false discovery rate controlling procedure for dependent test statistics. *Journal of Statistical Planning and Inference, 82,* 171–196.

Zhou, Y., Chunlei, Ke, Jiang, Q., Shahin, S., & Snapinn, S. (2015). Choosing appropriate metrics to evaluate adverse events in safety evaluation. *Therapeutic Innovation and Regulatory Science, 49,* 398–404.

Zink, R. C., Wolfinger, R. D., & Mann, G. (2013). Summarizing the incidence of adverse events using volcano plots and time windows. *Clinical Trials, 10,* 398–406.

Zink, R. C., Shen, L., Wolfinger, R. D., & Showalter, H. D. H. (2015). Assessment of methods to identify patient subgroups with enhanced treatment response in randomized clinical trials. In Z. Chen, A. Liu, Y. Qu, L. Tang, N. Ting, & Y. Tsong (Eds.), *Applied statistics in biomedicine and clinical trials design: Selected papers from 2013 ICSA/ISBS joint statistical meetings*. Cham, Switzerland: Springer.

Chapter 8
Meta-analysis for Rare Events in Clinical Trials

Ding-Geng Chen and Karl E. Peace

8.1 Introduction

Meta-analysis (MA) in clinical trials is defined as a statistical procedure for systematically combining data from multiple trials to reach a single conclusion with greater statistical power than any of the component trials. Specifically, the purpose of a meta-analysis is to combine the estimates of study effect-sizes using well-established approaches of the fixed-effects and random-effects models, which differ fundamentally on whether between-study heterogeneity is incorporated. These traditional meta-analysis models have played an increasingly important role in biopharmaceutical and medical sciences, and meta-analyses of clinical trial data have led to numerous scientific discoveries.

This chapter is organized to present an overview of methods and then an illustration of the methods. Section 8.2 gives an overview of meta-analysis using summary statistics using fixed-effects and random-effects models along with the quantification of heterogeneity with Q-statistic, τ^2 index, H index and I^2 index with emphasizing on binary data. Section 8.3 then illustrates the potential problems when these methods are used for clinical trials with rare events using the well-known Rosiglitazone meta-analysis data. We then introduce the R package `gmeta` for meta-analysis of rare events. Detailed analysis using open source R software is illustrated in this section so readers can follow the meta-analysis for their own data analysis.

D.-G. Chen (✉)
School of Social Work & Gillings School of Global Public Health,
University of North Carolina, Chapel Hill, NC 27599, USA
e-mail: dinchen@email.unc.edu

D.-G. Chen
University of Pretoria, Pretoria, South Africa

K. E. Peace
Jiann-Ping Hsu College of Public Health, Georgia Southern University,
Statesboro, GA 30458, USA

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_8

Section 8.4 concludes with a summary. Most of the materials of this chapter are summarized from Chen and Peace (2013) and the readers can read this chapter in connection with the book for greater understanding.

8.2 Overview of Meta-analysis

We encourage readers to first read Chap. 2 of Chen and Peace (2013) before embarking on a particular meta-analysis.

8.2.1 Summary Statistics and the Sources of Variations

In a typical meta-analysis with summary statistics, K independent studies are obtained to estimate a parameter of interest, such as the effect-size (ES) of efficacy β_k $(k = 1, 2, \ldots, K)$ between new treatment and a control. This analytic approach can be applied to a broad range of study designs, including single-arm or multiple-arm studies, randomized controlled trials, and observational studies. For ease of illustration, this discussion focuses on MA with two-arm studies, where β_k is some form of the effect-size between the two groups. The most popular choices for β_k for a continuous outcome are the mean difference or the standardized mean difference; for a dichotomous outcome, the most popular choices are odds ratio, risk ratio, and risk difference. In most cases, an estimated $\widehat{\beta}_k$ of the true β_k and its associated standard error could be directly extracted from each study. The ultimate goal of meta-analysis is to produce an optimal estimate of the overall population effect-size by pooling the estimates $\widehat{\beta}_k$ $(k = 1, 2, \ldots, K)$ from individual studies using appropriate statistical models.

Two sources of errors or variations exist in these summary statistics of $\widehat{\beta}_k$ from different studies with one source being the within-study variation and the other source the between-study variation. Within-study variation is caused by sampling error, which is random or non-systematic within each study. In contrast, between-study variation may result from systematic differences among studies. If the between-study variation can be verified as being zero, the effect estimates $\widehat{\beta}_k$ are considered homogeneous; Otherwise, the effect estimates are heterogeneous. In MA, the assumption of homogeneity states that β_k $(k = 1, 2, \ldots, K)$ is the same in all studies, that is

$$\beta_1 = \beta_2 = \cdots = \beta_K = \beta. \tag{8.1}$$

If these studies are homogeneous, then two commonly used meta-analysis models can be used: the fixed-effects MA model and random-effects MA model.

8.2.2 Effect-Size Calculations for Binary Data

8.2.2.1 Data Structure

For illustration purpose, let's detail the effect-size (ES) calculation with binary data for a clear understanding of meta-analysis models in the later sections.

Binary data are typically presented in a 2 by 2 table which is usually used to report the number of events and non-events in experimental treatment group (i.e., E) versus the control group (i.e., C).

From a total of K studies, the data from the kth study ($k = 1, \ldots, K$) can be represented as cells $x_{Ek}, n_{Ek} - x_{Ek}, x_{Ck}, n_{Ck} - x_{Ck}$ as shown in Table 8.1.

Commonly used ESs for binary data are the risk ratio, the risk-difference, and the odds-ratio. To simplify the notations, let's drop the subscript k.

8.2.2.2 ES with Risk-Ratio

The effect size (i.e., β_k) for the risk-ratio (RR) of experimental treatment the standard control is defined as:

$$ES = \frac{p_E}{p_C} = \frac{x_E/n_E}{x_C/n_C} \tag{8.2}$$

where p_E is the so-called risk (or risk probability) for the experimental treatment (E) which is computed as the total number of events (x_E) divided by the total number of patients (n_E), i.e., $p_E = \frac{x_E}{n_E}$ with similar notations for control (C).

To construct an approximate confidence interval based on the normal distribution, ES is transformed using the natural logarithm and then employing the delta-method where $lnES = ln(ES)$. The variance for $lnES$ can be shown to be

$$Var_{lnES} = \frac{1}{x_E} - \frac{1}{n_E} + \frac{1}{x_C} - \frac{1}{n_C} \tag{8.3}$$

Therefore, the approximated standard error is

$$SE_{lnES} = \sqrt{Var(lnES)} \tag{8.4}$$

Table 8.1 Nomenclature for 2 × 2 table of outcome by treatment

	Events	Non-events	Total event
Experimental treatment	x_{Ek}	$n_{Ek} - x_{Ek}$	n_{Ek}
Control	x_{Ck}	$n_{Ck} - x_{Ck}$	n_{Ck}

With this, the 95% CI for $lnES$ can be expressed as

$$(lnES - 1.96 \times SE_{lnES}, lnES + 1.96 \times SE_{lnES}). \tag{8.5}$$

We then transform back to the original scale for risk-ratio(RR) as:

$$\begin{aligned} RR &= exp(lnES) \\ L_{RR} &= exp(lnES - 1.96 \times SE_{lnES}) \\ U_{RR} &= exp(lnES + 1.96 \times SE_{lnES}) \end{aligned}$$

8.2.3 ES with Risk-Difference

Even though the risk-ratio is the most commonly used in binomial data, the risk difference is an ES which is easily understandable. The definition of risk difference (RD) is simply the difference of the risks between two treatments as:

$$ES_{RD} = \hat{p}_E - \hat{p}_C = \frac{x_E}{n_E} - \frac{x_C}{n_C} \tag{8.6}$$

using notations from previous section.

The variance of ES_{RD} can be estimated as:

$$Var\,(ES_{RD}) = \frac{\hat{p}_E(1-\hat{p}_E)}{n_E} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_C} \tag{8.7}$$

and the standard error (SE) is then calculated as $SE_{ES_{RD}} = \sqrt{Var\,(ES_{RD})}$. With the point estimate from Eq. (8.6) and its variance in Eq. (8.7), we can frame the same procedures for statistical inference similar to the risk ratio.

8.2.3.1 ES with Odds-Ratio

The odds-ratio (OR) associated with an event is defined as the ratio of the odds of the event in one study group to the odds of the event in another study group. The odds of the event for the treatment group is

$$OddsE = \frac{p_E}{1-p_E} = \frac{x_E}{n_E - x_E} \tag{8.8}$$

Similarly the odds of the event for the control group is

$$OddsC = \frac{p_C}{1-p_C} = \frac{x_C}{n_C - x_C} \tag{8.9}$$

Then the odds-ratio (OR) of the treatment group to the control group is as follows:

$$OR = \frac{OddsE}{OddsC} \tag{8.10}$$

To approximate the normal distribution in using odds-ratios, we usually convert the odds-ratio to the log scale and estimate the log odds-ratio and its standard error and use these numbers to perform the meta-analysis. Then we transform the results back into the original metric.

With this direction, the log odds-ratio is

$$LogOR = ln(OR) \tag{8.11}$$

The approximate variance can be derived using the delta method to expand (via Taylor series) the log-odds for both treatment and control about their expected values and the variance can then be obtained as follows:

$$var(logOR) = \frac{1}{x_E} + \frac{1}{n_E - x_E} + \frac{1}{x_C} + \frac{1}{n_C - x_C} \tag{8.12}$$

Therefore the approximate standard error is:

$$SE_{logOR} = \sqrt{V_{logOR}} \tag{8.13}$$

With these calculations in the log-scale, we then transform them back to original scale for odds-ratios (OR) using

$$OR = exp(logOR) \tag{8.14}$$

$$LL_{OR} = exp(LL_{logOR}) \tag{8.15}$$

and

$$UL_{OR} = exp(LL_{logOR}) \tag{8.16}$$

where *LL* and *UL* represent the lower and upper limits, respectively.

8.2.4 Fixed-Effects Meta-analysis

As shown in Eq. (8.1), a fixed-effects meta-analysis assumes homogeneity when the underlying population effect-sizes β_k are constant across all studies. A typical fixed-effects model is described as

$$\widehat{\beta_k} = \beta + \epsilon_k;\ k = 1, 2, \ldots, K, \tag{8.17}$$

where $\widehat{\beta_k}$ represents the effect-size for study k and β is the overall global population effect-size. The ϵ_k are the sampling error with mean 0 and KNOWN variance $\widehat{\sigma}^2_{\beta_k}$ which can be extracted or calculated from the individual studies. In general, the ϵ_k is assumed to follow a normal distribution $N(0, \widehat{\sigma}^2_{\beta_k})$. A pooled estimate of β in fixed-effects MA models is given by the weighted least squares estimation

$$\hat{\beta}_F = \frac{\Sigma_{k=1}^K w_k \widehat{\beta_k}}{\Sigma_{k=1}^K w_k}, \tag{8.18}$$

and the variance of $\hat{\beta}_F$ can be expressed as

$$Var\left(\hat{\beta}_F\right) = 1/\Sigma_{k=1}^K w_k \tag{8.19}$$

where a popular choice of weight is $w_k = 1/\widehat{\sigma}^2_{\beta_k}$ and variance $\widehat{\sigma}^2_{\beta_k}$ is estimated from study k. Hence, the 95% confidence interval of β_F is given by

$$\hat{\beta}_F - z_{0.025}\sqrt{Var\left(\hat{\beta}_F\right)} \le \beta \le \hat{\beta}_F + t_{0.025}\sqrt{Var\left(\widehat{\beta_F}\right)}, \tag{8.20}$$

where $z_{0.025}$ denotes the 2.5%-percentile of the standard normal distribution. Similarly, a statistical z-test can be formulated as:

$$z = \frac{\hat{\beta}_F - \beta}{\sqrt{Var\left(\hat{\beta}_F\right)}} \tag{8.21}$$

to be used to test $H_0 : \beta = 0$.

8.2.5 Random-Effects Meta-analysis

When meta-analyzing effect-sizes from a group of studies, it may be impractical to follow the assumption of the fixed-effects model that the K true effect-sizes are the same for all studies. When attempting to synthesize a group of studies using meta-analysis, we assume that the studies have enough in common to allow the data from those studies to be combined for statistical inference; however, it would be impractical to require that all studies in the group have identical true effect-sizes. It is impossible for independent studies to be identical in every respect. Therefore, heterogeneity is likely to exist in many meta-analyses. The model that takes heterogeneity into account is the following random-effects meta-analysis models:

$$\widehat{\beta}_k = \beta + b_k + \epsilon_k, k = 1, 2, \ldots, K, \tag{8.22}$$

where for study k, $\widehat{\beta}_k$ represents the observed effect-size and β the global population effect-size. b_k is now the random-effects with mean 0 and variance τ^2 representing the between-study heterogeneity, and ϵ_k is the sampling error with mean 0 and variance $\widehat{\sigma}^2_{\beta_k}$. It is assumed that b_k and ϵ_k are independent and follow normal distributions $N(0, \tau^2)$ and $N(0, \widehat{\sigma}^2_{\beta_k})$, respectively. Let $\beta_k = \beta + b_k$, $k = 1, 2, \ldots, K$. Then the random-effects model (8.22) can be simplified as

$$\widehat{\beta}_k = \beta_k + \epsilon_k, \tag{8.23}$$

where β_k represents the true effect-size for study k. All β_k $(k = 1, 2, \ldots, K)$ are random samples from the same normal population

$$\beta_k \sim N(\beta, \tau^2) \tag{8.24}$$

rather than being a constant in the fixed-effects MA in Eq. (2.2). Further, the marginal variance of $\widehat{\beta}_k$ is given by

$$Var\left(\widehat{\beta}_k\right) = \tau^2 + \widehat{\sigma}^2_{\beta_k}, \tag{8.25}$$

which is composed of two sources of variation, i.e., the between-study variance τ^2 and within-study variance $\widehat{\sigma}^2_{\beta_k}$. If the between-study variance $\tau^2 = 0$, the random-effects MA models (8.22) would reduce to the fixed-effects MA models (8.17).

Similar to the fixed-effects MA models, the within-study variance $\widehat{\sigma}^2_{\beta_k}$ can be obtained or calculated from each study k $(k = 1, 2, \ldots, K)$. However, the information for determining between-study variance τ^2 is often not available, and the methods commonly used for assessing between-study heterogeneity include the DerSimonian-Laird's method of moments (MM) in DerSimonian and Laird (1986), the maximum likelihood estimation (MLE) method in Hardy and Thompson (1998), and the restricted maximum likelihood (REML) method in Raudenbush and Bryk (1985). As the most commonly used estimator, MM is a distribution-free and non-iterative approach whereas both MLE and REML are parametric methods and need iteration for estimating τ^2.

The Method-of-moments (MM) uses the Q-statistic for testing the assumption of homogeneity:

$$Q = \sum_{k=1}^{K} w_k \left(\widehat{\beta}_k - \hat{\beta}_F\right)^2 \tag{8.26}$$

where $w_k = 1/\widehat{\sigma}^2_{\beta_k}$ is the weight from the kth study, $\widehat{\beta}_k$ is the kth study effect-size, and $\hat{\beta}_F$ is the global overall effect estimated from Eq. (8.18). It can be seen that Q is calculated as: (1) compute the deviations of each effect-size from the meta-estimate and square them (i.e., $(\hat{\beta}_k - \hat{\beta}_F)^2$), (2) weight these values by the inverse-variance

for each study, and (3) then sum these values across all K studies to produce a weighted sum of squares (WSS) to obtain the heterogeneity measure Q.

From Eq. (8.26), it can be shown that the expected value of Q is

$$E(Q) = (K-1) + U \times \tau^2 \tag{8.27}$$

where $U = \sum_{k=1}^{K} w_k - \frac{\sum_{k=1}^{K} w_k^2}{\sum_{k=1}^{K} w_k}$.

Under the assumption of no heterogeneity (all studies have the same effect-size), then τ^2 would be zero and $E(Q) = df = K - 1$. Based on this heterogeneity measure Q, the test of heterogeneity is conducted to address the null hypothesis that the effect-sizes β_k from all studies share a common effect-size β (i.e., the assumption of homogeneity) and then test this hypothesis where the test statistic is constructed using Q as a central χ^2 distribution with degrees of freedom of $df = K - 1$. However, a cautionary note is warranted: This χ^2-test using the Q-statistic has poor statistical power to detect true heterogeneity for a meta-analysis with a small number of studies K, but excessive power to detect negligible variability with a large number of studies as discussed in Harwell (1997) and Hardy and Thompson (1998). Thus, a nonsignificant test using Q-statistic from a small number of studies can lead to an erroneous selection of a fixed-effects model when possible true heterogeneity exists among the studies, and vice versa. The inability to conclude statistically significant heterogeneity in a meta-analysis of a small number of studies at the 0.05 level of significance is similar to failing to detect statistically significant treatment-by-center interaction in an MRCT. In these settings, many analysts might choose to conduct the test of homogeneity at the 0.10 level as a means of increasing the power of the test.

From Eq. (8.27), the method-of-moments (MM) estimate of τ^2 can be shown as follows:

$$\widehat{\tau}^2 = \max\left(0, \frac{Q-(K-1)}{U}\right) \tag{8.28}$$

The truncation at zero in (8.28) ensures the variance estimate is non-negative.

Therefore, the estimate of the global population effect-size in random-effects MA is given by

$$\widehat{\beta}_R = \frac{\Sigma_{k=1}^{K} w_k^* \widehat{\beta}_k}{\Sigma_{k=1}^{K} w_k^*} \tag{8.29}$$

where $w_k^* = 1/\left(\widehat{\sigma}_{\beta_k}^2 + \widehat{\tau}^2\right)$. The variance of $\widehat{\beta}_R$ is simply

$$Var\left(\widehat{\beta}_R\right) = 1/\sum_{k=1}^{K} w_k^*$$

and the 95% confidence interval can be constructed by $\widehat{\beta}_R - z_{0.025}\sqrt{Var\left(\widehat{\beta}_R\right)} \leq \beta \leq \widehat{\beta}_R + z_{0.025}\sqrt{Var\left(\widehat{\beta}_R\right)}$.

The statistical test can be similarly formulated by:

$$z = \frac{\hat{\beta}_R - \beta}{\sqrt{Var\left(\hat{\beta}_R\right)}} \tag{8.30}$$

to be used to test $H_0 : \beta = 0$ in the random-effects MA framework.

8.3 Meta-analysis with Rare-Events

8.3.1 Potential Problems

All the methods presented thus far for meta-analysis in Sect. 8.2 are based on large sample theory as well as the theory of large sample approximations. For rare events, these methods usually break down.

For example, when events are zeros, the methods for risk-ratio and odds-ratio cannot be used and when the events are rare, but not all zeros, the variance estimates for these methods are not robust which may lead to unreliable statistical inferences. The typical remedies are to remove the studies with zero events from the meta analysis, or add a small value, say 0.5, to the rare events which could lead to biased statistical inferences as pointed out by Tian et al. (2009) and Cai et al. (2010).

In this chapter, we use the well-known Rosiglitazone meta-analysis data to illustrate the bias when classical meta-analysis methods are used for rare events.

8.3.2 The Rosiglitazone Meta-analysis

In a meta-analysis for the effect of rosiglitazone on the risk of myocardial infarction (MI) and death from cardiovascular causes, Nissen and Wolski (2007) searched the available published literature and found 116 potentially relevant studies where 42 of these met the inclusion criteria. Data were then extracted from the 42 publications and combined using a fixed-effects meta-analysis model. This yielded an odds-ratio for the rosiglitazone group to the control group of 1.43 with 95% CI of (1.03, 1.98) and p-value = 0.03 for MI; and 1.64 with 95% CI of (0.98, 2.74) and p-value = 0.06 for death from cardiovascular causes. Based on these results, the authors concluded that rosiglitazone use was statistically significantly associated with risk of myocardial infarction, and was borderline statistically significant with death from cardiovascular causes. Therefore using rosiglitazone for the treatment of Type-2 diabetes could lead to serious adverse cardiovascular effects.

Since its publication, numerous authors questioned the validity of the analysis and interpretation of the results. For example, Shuster and Schatz (2008)

(which is online available at http://care.diabetesjournals.org/content/31/3/e10.full.pdf) pointed out that the fixed-effects meta-analysis was inappropriate. They reanalyzed 48 (not 42) eligible studies via a new random-effects method (Shuster et al. 2007) that yielded different conclusions; i.e., a strong association with cardiac death was found, but there was no significant association with myocardial infarction. Other meta-analyses of data from the studies can be found from Dahabreh (2008), Tian et al. (2009), Cai et al. (2010) and Lane (2012) (online publication available at http://www.ncbi.nlm.nih.gov/pubmed/22218366).

In this section, we further illustrate meta-analysis for rare events using this data with R implementations in `gmeta`.

8.3.3 Step-by-Step Data Analysis in R

8.3.3.1 Load the Data

The data from Tian et al. (2009), which is available as the supplementary material at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2648899/bin/kxn034_index.html. This data can be loaded into R with the following R code chunk:

```
> # Load the Rosiglitazone data from excel file
> require(gdata)
> # Print the first 6 studies
> head(dat)

ID  Study  n.TRT MI.TRT Death.TRT n.CTRL MI.CTRL Death.CTRL
1 49653/011  357    2       1     176       0     0
2 49653/020  391    2       0     207       1     0
3 49653/024  774    1       0     185       1     0
4 49653/093  213    0       0     109       1     0
5 49653/094  232    1       1     116       0     0
6    100684   43    0       0      47       1     0
```

With this dataframe, we perform meta-analyses of both the risk difference (RD) and odds-ratio (OR) for myocardial infarction (MI) and cardiovascular death (Death). We contrast the results from the classical fixed-effects and random-effects models using the R package `meta` to the results from the *confidence distribution* (CD) implemented in the R package `gmeta`.

8.3.3.2 Meta-analysis for Myocardial Infarction (MI)

To analyze the data for MI, we first create a dataframe (only for MI) as follows:

```
> datMI    = dat[,c("MI.TRT","MI.CTRL","n.TRT","n.CTRL")]
```

For classical fixed-effects and random-effects meta-analysis, we make use of the R library `meta`, introduced in previous chapters, and use the inverse weighting method to combine studies. This is implemented in the following R code chunk:

```
> # Load the library
> library(meta)
> # Call metabin with RD=risk difference
> MI.RD.wo = metabin(MI.TRT,n.TRT,MI.CTRL,n.CTRL,data=datMI,
    incr=0, method="Inverse", sm="RD")
> # Print the summary
> summary(MI.RD.wo)

Number of studies combined: k=48

                         RD           95%-CI     z  p.value
Fixed effect model    0.002  [0.001; 0.003] 3.24   0.0012
Random effects model 0.002  [0.001; 0.003] 3.24   0.0012

Quantifying heterogeneity:
tau^2 < 0.0001; H = 1 [1; 1]; I^2 = 0% [0%; 0%]

Test of heterogeneity:
    Q d.f.  p.value
 27.9   47   0.9879

Details on meta-analytical method:
- Inverse variance method
- DerSimonian-Laird estimator for tau^2
```

As seen from the summary, the combined RD = 0.0018 with 95% CI of $(7e-04, 0.0028)$ and a p-value = 0.0012 for both fixed-effects and random-effects models—since the `Test of heterogeneity` is not statistically significant (p-value = 0.9879 and $\hat{\tau}^2 \approx 0$). Even though the RD is small and the left endpoint of the CI is just to the right of 0, these results are consistent with the conclusion that MIs in rosiglitazone group are statistically significantly higher than in the control group.

Note that in the above R code chunk, the option `incr` is set to zero which means no value is added to the zero MIs. In this dataframe, there are 10 studies with zero MIs for both rosiglitazone and control. The standard errors for the RD corresponding to these studies cannot be computed which is set to zero as default in this R function call.

A typical way to adjust the zero MIs is to add a small increment of 0.5 to them as a correction for lack of continuity, which is the default setting in the R function call to `metabin` as follows:

```
> # Call metabin with default setting to add 0.5
> MI.RD = metabin(MI.TRT,n.TRT,MI.CTRL,n.CTRL,data=datMI,
          method="Inverse", sm="RD")
> # Print the summary
> summary(MI.RD)

Number of studies combined: k=48

                          RD      95
Fixed effect model    0.001  [0; 0.003] 1.73   0.0834
Random effects model 0.001  [0; 0.003] 1.73   0.0834

Quantifying heterogeneity:
tau^2 < 0.0001; H = 1 [1; 1]; I^2 = 0

Test of heterogeneity:
     Q d.f. p.value
 17.98   47       1

Details on meta-analytical method:
- Inverse variance method
- DerSimonian-Laird estimator for tau^2
```

With 0.5 added to the zero cells, we see from the output that the combined RD is now 0.0014 with 95% CI of $(-2e-04, 0.0029)$ and p-value = 0.0834 for both fixed-effects and random-effects models. The conclusion changed from statistically significant to statistically non-significant. Readers may want to try to add different increments to the zero cells and examine the effects of this artificial correction (although well founded in history of the analysis of contingency table data) for lack of continuity. In fact, Sweeting et al. (2004) provided compelling evidence that imputing arbitrary numbers to zero cells in continuity correction can result in very different conclusions.

Tian et al. (2009) developed an exact and efficient inference procedure to use all the data without this artificial continuity correction. This is a special case of the *confidence distribution* (CD) framework as proved in the **Supplementary Notes** at http://stat.rutgers.edu/home/gyang/researches/gmetaRpackage/. This method is implemented into `gmeta` as `method`="`exact2`". The R code to implement this method is as follows:

```
> # Call "gmeta" with method="exact2"
> MI.exactTianRD =  gmeta(datMI,gmi.type="2x2",method="exact2",
          ci.level=0.95,n=2000)
```

The summary of this modeling can be printed as follows:

```
> summary(MI.exactTianRD)

    Exact Meta-Analysis Approach through CD-Framework
Call:
gmeta.default(gmi = datMI, gmi.type = "2x2", method = "exact2",
    n = 2000, ci.level = 0.95)

Combined CD Summary:
                mean    median  stddev     CI.1     CI.2
exp1        -4.53e-03 -5.81e-03 0.00619 -0.01777 0.020567
exp2         6.12e-04 -1.16e-03 0.00878 -0.01396 0.018077
exp3         5.97e-03  3.73e-03 0.00727 -0.00565 0.025406
exp4         1.12e-02  1.05e-02 0.01256 -0.01489 0.044330
exp5        -2.44e-03 -4.30e-03 0.00819 -0.01949 0.031133
exp6         1.75e-02  1.95e-02 0.03239       NA       NA
exp7        -7.89e-03 -7.47e-03 0.01245 -0.03842 0.029153
exp8        -2.24e-02 -3.27e-02 0.02650       NA 0.027238
exp9        -2.50e-03 -2.56e-03 0.00389 -0.01194 0.009509
exp10       -1.84e-03 -4.05e-03 0.00694 -0.01605 0.026811
exp11        3.49e-03  3.79e-03 0.00504 -0.01164 0.016145
exp12       -1.53e-03 -3.44e-03 0.00622 -0.01209 0.017555
exp13       -3.08e-03 -5.39e-03 0.01073 -0.02021 0.012985
exp14       -3.91e-03 -4.61e-03 0.00536 -0.01519 0.017104
exp15        1.00e-03 -1.08e-03 0.00861 -0.01378 0.019149
exp16        5.87e-03  5.62e-03 0.01363 -0.01808 0.039631
exp17        9.03e-03  6.97e-03 0.02226 -0.03358 0.055565
exp18       -7.81e-03 -9.18e-03 0.01055 -0.02966 0.033602
exp19       -1.35e-02 -1.61e-02 0.01769 -0.05159 0.025194
exp20        2.48e-03 -3.53e-04 0.00951 -0.01477 0.030524
exp21        8.63e-03  7.78e-03 0.01272 -0.02793 0.040218
exp22        6.09e-03  5.53e-03 0.00878 -0.01952 0.028042
exp23       -1.46e-02 -1.71e-02 0.02632       NA       NA
exp24       -1.49e-02 -2.85e-02 0.03846       NA 0.049259
exp25        8.28e-03  7.01e-03 0.00615 -0.00254 0.023689
exp26        7.00e-03  5.72e-03 0.02451 -0.04541       NA
exp27       -6.34e-03 -7.63e-03 0.01003 -0.03105 0.025329
exp28       -4.22e-03 -4.17e-03 0.00649 -0.01995 0.015046
exp29       -1.03e-02 -1.18e-02 0.01668 -0.05235 0.040833
exp30       -5.72e-03 -5.40e-03 0.00893 -0.02750 0.021104
exp31        2.79e-03 -1.43e-06 0.01502 -0.02461 0.047615
exp32       -9.28e-05 -8.58e-04 0.00241 -0.00421 0.009685
exp33        8.12e-04 -8.25e-05 0.00287 -0.00417 0.009115
exp34        5.67e-03  3.73e-03 0.01191 -0.01673 0.030232
exp35       -3.27e-03 -3.84e-03 0.00512 -0.01577 0.013017
exp36       -3.90e-03 -4.15e-03 0.00592 -0.01818 0.013397
exp37       -1.72e-03 -3.43e-03 0.00589 -0.01445 0.023542
exp38        1.56e-04 -1.94e-04 0.00651 -0.01712 0.018428
exp39        6.13e-04 -2.07e-03 0.00806 -0.01238 0.024941
exp40       -2.41e-04 -2.33e-03 0.00715 -0.01234 0.021490
exp41       -2.39e-03 -2.52e-03 0.00200 -0.00651 0.001540
```

```
exp42       -4.70e-03 -4.70e-03 0.00445 -0.01419 0.003493
exp43        2.94e-03 -6.03e-07 0.01802 -0.02813 0.056682
exp44        2.10e-03 -1.27e-04 0.00812 -0.01255 0.025546
exp45       -3.43e-04 -5.37e-04 0.01453 -0.03956 0.038902
exp46       -8.29e-05 -4.24e-03 0.04255       NA       NA
exp47        1.16e-04 -5.10e-07 0.00532 -0.01408 0.015145
exp48        1.43e-03 -1.07e-05 0.00424 -0.00507 0.013882
combined.cd -1.77e-03 -2.21e-03 0.00188 -0.00386 0.000878

Confidence level= 0.95
```

The last row contains the combined estimates and can be produced as follows:

```
> summary(MI.exactTianRD)$mms[49,]

                mean   median  stddev     CI.1     CI.2
combined.cd -0.00177 -0.00221 0.00188 -0.00386 0.000878
```

We see that the mean difference is -0.00177 with 95% CI of $(-0.00386, 0.00088)$ indicating no statistically significantly difference between rosiglitazone group and the control group on MI.

We now analyze the MI dataframe using the odds ratio. Similarly, the classical fixed-effects and random-effects models can be implemented as follows:

```
> # Call metabin without 0.5 correction
> MI.OR.wo = metabin(MI.TRT,n.TRT,MI.CTRL,n.CTRL,data=datMI,
         incr=0,method="Inverse", sm="OR")
> # Summary
> summary(MI.OR.wo)

Number of studies combined: k=38

                        OR          95
Fixed effect model   1.29  [0.895; 1.85] 1.36   0.1736
Random effects model 1.29  [0.895; 1.85] 1.36   0.1736

Quantifying heterogeneity:
tau^2 < 0.0001; H = 1 [1; 1]; I^2 = 0

Test of heterogeneity:
   Q d.f. p.value
 5.7   37       1

Details on meta-analytical method:
- Inverse variance method
- DerSimonian-Laird estimator for tau^2


> # Call metabin with default 0.5 correction
> MI.OR = metabin(MI.TRT,n.TRT,MI.CTRL,n.CTRL,data=datMI,
             method="Inverse", sm="OR")
> # Print the Summary
> summary(MI.OR)
```

```
Number of studies combined: k=38

                         OR        95
Fixed effect model   1.29  [0.94; 1.76] 1.57   0.1161
Random effects model 1.29  [0.94; 1.76] 1.57   0.1161

Quantifying heterogeneity:
tau^2 < 0.0001; H = 1 [1; 1]; I^2 = 0

Test of heterogeneity:
     Q d.f.  p.value
 16.22   37   0.9988

Details on meta-analytical method:
- Inverse variance method
- DerSimonian-Laird estimator for tau^2
```

We see that with or without the default 0.5 continuity correction, the 95% CIs and p-values are slightly different, but yield the same conclusion that there is no statistically significantly difference between the rosiglitazone group and the control group on MI.

We now can call `gmeta` for the exact method using the odds ratio, which is implemented as follows:

```
> # Call "gmeta" for "exact1" on OR
> MI.exactLiuOR = gmeta(datMI,gmi.type="2x2",
         method="exact1", ci.level=0.95,n=2000)
> # Print the summary
> summary(MI.exactLiuOR)

    Exact Meta-Analysis Approach through CD-Framework
Call:
gmeta.default(gmi = datMI, gmi.type = "2x2", method = "exact1",
    n = 2000, ci.level = 0.95)

Combined CD Summary:
               mean  median stddev   CI.1  CI.2
exp1            Inf      NA    Inf -1.951   Inf
exp2         0.1141 -0.0044  1.360 -2.525 3.446
exp3        -1.4316 -1.4333  1.631 -5.110 2.233
exp4           -Inf      NA    Inf   -Inf 2.274
exp5            Inf      NA    Inf -3.636   Inf
exp6           -Inf      NA    Inf   -Inf 3.033
exp7            Inf      NA    Inf -2.920   Inf
exp8         0.9942  0.9346  0.888 -0.669 3.002
exp9            Inf      NA    Inf -2.939   Inf
exp10           Inf      NA    Inf -3.687   Inf
exp11          -Inf      NA    Inf   -Inf 2.971
exp12           Inf      NA    Inf -2.625   Inf
exp13        0.7556  0.6374  1.360 -1.882 4.088
exp14           Inf      NA    Inf -1.922   Inf
exp15        0.0593 -0.0592  1.360 -2.581 3.392
```

```
exp16        -0.6514 -0.6520  1.634 -4.320 3.018
exp17        -0.8065 -0.6886  1.366 -4.143 1.840
exp18            Inf      NA    Inf -1.928   Inf
exp19         1.1847  1.0326  1.266 -1.135 4.398
exp20            NaN      NA    Inf   -Inf   Inf
exp21           -Inf      NA    Inf   -Inf 2.928
exp22           -Inf      NA    Inf   -Inf 2.933
exp23            Inf      NA    Inf -2.909   Inf
exp24            Inf      NA    Inf -2.970   Inf
exp25           -Inf      NA    Inf   -Inf 0.534
exp26        -0.4690 -0.4347  0.978 -2.603 1.466
exp27            Inf      NA    Inf -2.979   Inf
exp28            Inf      NA    Inf -2.898   Inf
exp29            Inf      NA    Inf -2.956   Inf
exp30            Inf      NA    Inf -2.921   Inf
exp31            NaN      NA    Inf   -Inf   Inf
exp32            Inf      NA    Inf -4.084   Inf
exp33            NaN      NA    Inf   -Inf   Inf
exp34        -0.8514 -0.7333  1.362 -4.183 1.788
exp35            Inf      NA    Inf -2.973   Inf
exp36            Inf      NA    Inf -2.876   Inf
exp37            Inf      NA    Inf -3.656   Inf
exp38            NaN      NA    Inf   -Inf   Inf
exp39            Inf      NA    Inf -4.306   Inf
exp40            Inf      NA    Inf -4.107   Inf
exp41         0.5133  0.5055  0.428 -0.314 1.385
exp42         0.2739  0.2760  0.251 -0.226 0.762
exp43            NaN      NA    Inf   -Inf   Inf
exp44            NaN      NA    Inf   -Inf   Inf
exp45            NaN      NA    Inf   -Inf   Inf
exp46            NaN      NA    Inf   -Inf   Inf
exp47            NaN      NA    Inf   -Inf   Inf
exp48            NaN      NA    Inf   -Inf   Inf
combined.cd   0.3300  0.3301  0.184 -0.028 0.694

Confidence level= 0.95
```

The combined results from this summary are on the log scale, and we transform back to the OR as follows:

```
> # Use `exp' function to transform back
> exp(summary(MI.exactLiuOR)$mms[49,])

            mean median stddev  CI.1 CI.2
combined.cd 1.39   1.39    1.2 0.972    2
```

This give the OR of 1.39 with 95% CI of (0.972, 2) which again indicates that there is no statistically significantly difference between the rosiglitazone group and the control group on MI.

We summarize the analyses using the novel *confidence distributions* approach implemented in `gmeta` in Fig. 8.1 with the following R code chunk where

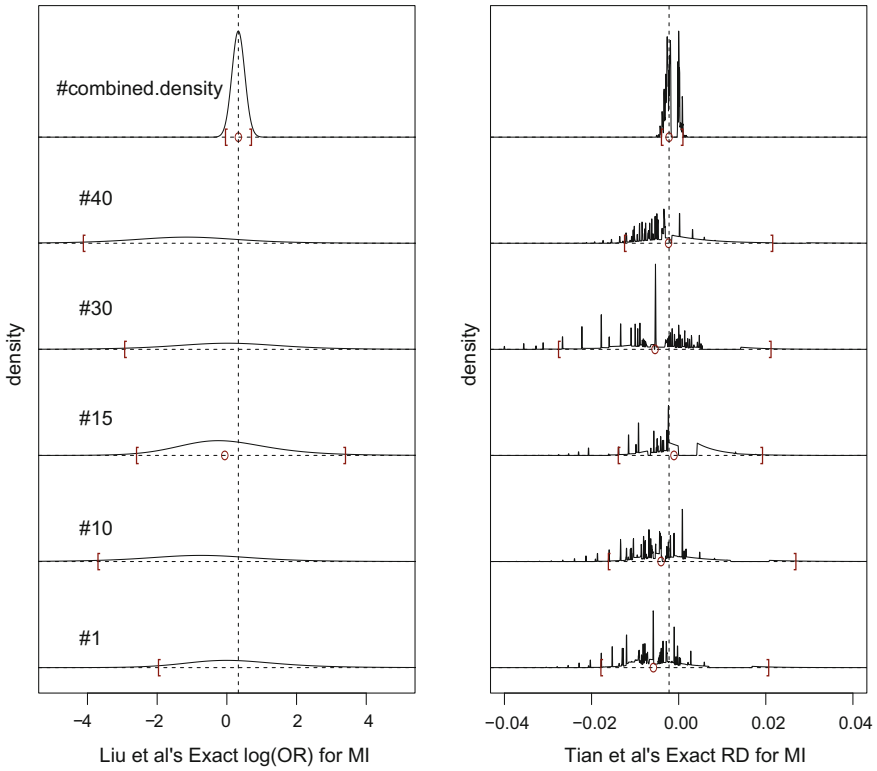


Fig. 8.1 Confidence Distributions from Both Exact Methods

we only include the CDs for studies 1, 10, 15, 30, 40 as well as the combined confidence distribution:

```
> # Plot the gmeta confidence distributions
> par(mfrow=c(1,2))
> plot(MI.exactLiuOR,  trials=c(1,10,15,30,40), option=T,
     xlim=c(-5,5),xlab="Liu et al's Exact log(OR) for MI")
> plot(MI.exactTianRD, trials=c(1,10,15,30,40), option=T,
     xlim=c(-0.04,0.04), xlab="Tian et al's Exact RD for MI")
```

8.3.3.3 Data Analysis for Cardiovascular Death (Death)

Similarly we use the same steps to analyze the data for cardiovascular death (Death). We first create a dataframe only for Death as follows:

```
> datDeath   = dat[,c("Death.TRT","Death.CTRL","n.TRT","n.CTRL")]
```

For risk difference, the classical fixed-effects and random-effects meta-analysis can be performed using the following R code chunk:

```
> # Call metabin with RD=risk difference
> Death.RD.wo = metabin(Death.TRT,n.TRT,Death.CTRL,n.CTRL,
     data=datDeath,incr=0, method="Inverse", sm="RD")
> # Print the summary
> summary(Death.RD.wo)

Number of studies combined: k=48

                         RD      95
Fixed effect model   0.001  [0; 0.002] 2.6   0.0094
Random effects model 0.001  [0; 0.002] 2.6   0.0094

Quantifying heterogeneity:
tau^2 < 0.0001; H = 1 [1; 1]; I^2 = 0

Test of heterogeneity:
     Q d.f. p.value
 13.69   47       1

Details on meta-analytical method:
- Inverse variance method
- DerSimonian-Laird estimator for tau^2

> # Call metabin with default setting to add 0.5
> Death.RD = metabin(Death.TRT,n.TRT,Death.CTRL,n.CTRL,
     data=datDeath, method="Inverse", sm="RD")
> # Print the summary
> summary(Death.RD)

Number of studies combined: k=48

                         RD         95
Fixed effect model   0.001 [-0.001;0.002] 0.943 0.3455
Random effects model 0.001 [-0.001;0.002] 0.943 0.3455

Quantifying heterogeneity:
tau^2 < 0.0001; H = 1 [1; 1]; I^2 = 0

Test of heterogeneity:
    Q d.f. p.value
 7.92   47       1

Details on meta-analytical method:
- Inverse variance method
- DerSimonian-Laird estimator for tau^2
```

Again, we see from the summaries that the combined RD = 0.001 with 95% CI of (0, 0.002) and a p-value = 0.0094 for both fixed-effects and random-effects models without continuity correction. This statistical significance vanishes when 0.5 is added to the zero cells in 25 studies. The combined RD is now 0.001 with 95% CI

of $(-0.001, 0.002)$ and a p-value = 0.943 for both fixed-effects and random-effects models.

With `gmeta` the risk difference is implemented as follows:

```
> # Call "gmeta" with method="exact2"
> Death.exactTianRD =  gmeta(datDeath,gmi.type="2x2",
         method="exact2", ci.level=0.95,n=2000)
```

The summary for this modeling is printed as follows:

```
> summary(Death.exactTianRD)

    Exact Meta-Analysis Approach through CD-Framework

Call:
gmeta.default(gmi = datDeath, gmi.type = "2x2", method = "exact2",
    n = 2000, ci.level = 0.95)

Combined CD Summary:
                mean    median   stddev     CI.1    CI.2
exp1        -1.55e-03 -2.97e-03 0.005132 -0.01274 0.02063
exp2         1.08e-03 -1.15e-06 0.004528 -0.00749 0.01427
exp3         2.08e-03 -9.16e-07 0.005188 -0.00554 0.01723
exp4         1.81e-03 -4.80e-07 0.008317 -0.01362 0.02693
exp5        -2.94e-03 -4.31e-03 0.008163 -0.01947 0.03113
exp6         1.31e-04 -1.19e-03 0.023533       NA      NA
exp7         6.87e-05 -5.92e-07 0.008521 -0.02292 0.02371
exp8        -6.73e-03 -1.45e-02 0.026914       NA      NA
exp9         4.45e-05  3.23e-05 0.002701 -0.00718 0.00772
exp10        1.78e-03 -4.23e-04 0.006930 -0.01028 0.02180
exp11       -3.02e-03 -5.11e-03 0.010069 -0.01917 0.01200
exp12        2.52e-03 -3.20e-07 0.006694 -0.00725 0.02227
exp13        3.81e-03  4.10e-03 0.005403 -0.01231 0.01747
exp14        1.11e-03 -4.11e-07 0.004256 -0.00700 0.01394
exp15       -4.12e-03 -5.03e-03 0.005708 -0.01607 0.01830
exp16        4.84e-03  5.62e-03 0.013645 -0.01808      NA
exp17        2.48e-04 -1.00e-03 0.010064 -0.02675 0.02961
exp18       -3.44e-03 -4.51e-03 0.009007 -0.02128 0.03371
exp19       -6.44e-03 -7.02e-03 0.010694 -0.03353 0.02605
exp20       -3.97e-03 -5.54e-03 0.009487 -0.02297 0.03750
exp21        1.54e-04 -3.44e-04 0.008721 -0.02279 0.02449
exp22        1.47e-04 -1.65e-04 0.006004 -0.01583 0.01705
exp23        8.78e-05 -7.53e-07 0.018110       NA      NA
exp24       -3.63e-05 -1.75e-03 0.026734       NA      NA
exp25       -9.62e-04 -1.89e-03 0.003275 -0.00815 0.01320
exp26       -4.80e-03 -1.32e-02 0.025969       NA 0.03232
exp27       -1.21e-02 -1.46e-02 0.012116       NA 0.02518
exp28       -4.21e-03 -4.17e-03 0.006489 -0.01995 0.01505
exp29        1.58e-04 -3.29e-04 0.011882 -0.03109 0.03325
exp30       -5.62e-03 -5.39e-03 0.008944 -0.02750 0.02110
exp31        1.27e-03 -9.28e-07 0.015018 -0.02465      NA
exp32       -9.40e-05 -8.58e-04 0.002395 -0.00421 0.00968
exp33       -6.91e-04 -1.52e-03 0.002826 -0.00651 0.01123
```

```
exp34         5.64e-03  4.19e-03 0.008280 -0.01679 0.02624
exp35        -3.28e-03 -3.86e-03 0.005116 -0.01577 0.01302
exp36        -7.01e-05 -1.60e-04 0.003964 -0.01086 0.01088
exp37         1.57e-03 -1.90e-04 0.006005 -0.00930 0.01914
exp38         1.51e-04 -1.95e-04 0.006493 -0.01709 0.01843
exp39         3.30e-04 -2.08e-03 0.008065 -0.01239 0.02494
exp40        -6.02e-04 -2.33e-03 0.007125 -0.01233 0.02149
exp41        -8.61e-04 -9.91e-04 0.001884 -0.00499 0.00301
exp42         1.71e-04  1.26e-04 0.001499 -0.00377 0.00351
exp43         8.82e-04 -3.89e-07 0.018000 -0.02815       NA
exp44         1.87e-03 -1.20e-04 0.008131 -0.01253 0.02555
exp45        -3.66e-04 -5.37e-04 0.014531       NA       NA
exp46        -3.34e-04 -4.23e-03 0.042551       NA       NA
exp47         1.23e-04 -5.15e-07 0.005314 -0.01412 0.01515
exp48         1.43e-03 -6.97e-06 0.004238 -0.00507 0.01388
combined.cd -7.59e-04 -8.93e-04 0.000622 -0.00233 0.00135

Confidence level= 0.95
```

The last row contained the combined estimates and is produced as follows:

```
> summary(Death.exactTianRD)$mms[49,]

                 mean    median   stddev     CI.1    CI.2
combined.cd -0.000759 -0.000893 0.000622 -0.00233 0.00135
```

We see that the mean difference is −0.000759 with 95% CI of (−0.00233, 0.00135) indicating no statistically significant difference between rosiglitazone group and the control group on cardiovascular death.

Similarly for the odds ratio, the classical fixed-effects and random-effects models are implemented as follows:

```
> # Call metabin without 0.5 correction
> Death.OR.wo = metabin(Death.TRT,n.TRT,Death.CTRL,n.CTRL,
     data=datDeath,incr=0,method="Inverse", sm="OR")
> # Summary
> summary(Death.OR.wo)

Number of studies combined: k=23

                       OR         95%-CI     z  p.value
Fixed effect model   1.2  [0.642; 2.24] 0.568   0.5699
Random effects model 1.2  [0.642; 2.24] 0.568   0.5699

Quantifying heterogeneity:
tau^2 < 0.0001; H = 1 [1; 1]; I^2 = 0% [0%; 0%]

Test of heterogeneity:
    Q d.f. p.value
 1.02   22       1

Details on meta-analytical method:
- Inverse variance method
- DerSimonian-Laird estimator for tau^2
```

```
> # Call metabin with default 0.5 correction
> Death.OR = metabin(Death.TRT,n.TRT,Death.CTRL,n.CTRL,
     data=datDeath, method="Inverse", sm="OR")
> # Print the Summary
> summary(Death.OR)

Number of studies combined: k=23

                        OR         95%-CI    z  p.value
Fixed effect model   1.31  [0.805; 2.13] 1.08   0.2783
Random effects model 1.31  [0.805; 2.13] 1.08   0.2783

Quantifying heterogeneity:
tau^2 < 0.0001; H = 1 [1; 1]; I^2 = 0% [0%; 0%]

Test of heterogeneity:
    Q d.f. p.value
 4.79   22       1

Details on meta-analytical method:
- Inverse variance method
- DerSimonian-Laird estimator for tau^2
```

We see that with or without the default 0.5 continuity correction, the 95% CIs and p-values are slightly different, but yield the same conclusion that there is no statistically significant difference between the rosiglitazone group and the control group on cardiovascular death.

Now we call `gmeta` for the exact method for the odds ratio which is implemented as follows:

```
> # Call "gmeta" for "exact1" on OR
> Death.exactLiuOR  =  gmeta(datDeath,gmi.type="2x2",
         method="exact1", ci.level=0.95,n=2000)
> # Print the summary
> summary(Death.exactLiuOR)

        Exact Meta-Analysis Approach through CD-Framework

Call:
gmeta.default(gmi = datDeath, gmi.type = "2x2", method = "exact1",
    n = 2000, ci.level = 0.95)

Combined CD Summary:
             mean median stddev   CI.1 CI.2
exp1          Inf     NA    Inf -3.651  Inf
exp2          NaN     NA    Inf   -Inf  Inf
exp3          NaN     NA    Inf   -Inf  Inf
exp4          NaN     NA    Inf   -Inf  Inf
exp5          Inf     NA    Inf -3.636  Inf
exp6          NaN     NA    Inf   -Inf  Inf
exp7          NaN     NA    Inf   -Inf  Inf
exp8        0.461  0.426  0.979 -1.473 2.59
```

```
exp9           NaN     NA    Inf   -Inf  Inf
exp10          NaN     NA    Inf   -Inf  Inf
exp11        0.779  0.661  1.360 -1.859 4.11
exp12          NaN     NA    Inf   -Inf  Inf
exp13         -Inf     NA    Inf   -Inf 2.95
exp14          NaN     NA    Inf   -Inf  Inf
exp15          Inf     NA    Inf -1.934  Inf
exp16       -0.651 -0.652  1.634 -4.320 3.02
exp17          NaN     NA    Inf   -Inf  Inf
exp18          Inf     NA    Inf -3.627  Inf
exp19          Inf     NA    Inf -2.937  Inf
exp20          Inf     NA    Inf -3.657  Inf
exp21          NaN     NA    Inf   -Inf  Inf
exp22          NaN     NA    Inf   -Inf  Inf
exp23          NaN     NA    Inf   -Inf  Inf
exp24          NaN     NA    Inf   -Inf  Inf
exp25          Inf     NA    Inf -3.653  Inf
exp26        0.712  0.594  1.365 -1.934 4.05
exp27          Inf     NA    Inf -1.278  Inf
exp28          Inf     NA    Inf -2.898  Inf
exp29          NaN     NA    Inf   -Inf  Inf
exp30          Inf     NA    Inf -2.921  Inf
exp31          NaN     NA    Inf   -Inf  Inf
exp32          Inf     NA    Inf -4.084  Inf
exp33          Inf     NA    Inf -3.719  Inf
exp34         -Inf     NA    Inf   -Inf 2.85
exp35          Inf     NA    Inf -2.973  Inf
exp36          NaN     NA    Inf   -Inf  Inf
exp37          NaN     NA    Inf   -Inf  Inf
exp38          NaN     NA    Inf   -Inf  Inf
exp39          Inf     NA    Inf -4.306  Inf
exp40          Inf     NA    Inf -4.107  Inf
exp41        0.183  0.180  0.435 -0.672 1.05
exp42       -0.246 -0.186  0.877 -2.235 1.40
exp43          NaN     NA    Inf   -Inf  Inf
exp44          NaN     NA    Inf   -Inf  Inf
exp45          NaN     NA    Inf   -Inf  Inf
exp46          NaN     NA    Inf   -Inf  Inf
exp47          NaN     NA    Inf   -Inf  Inf
exp48          NaN     NA    Inf   -Inf  Inf
combined.cd  0.385  0.385  0.343 -0.268 1.09

Confidence level= 0.95
```

The combined results from this summary are on the log scale. We transform back to the OR as follows:

```
> exp(summary(Death.exactLiuOR)$mms[49,])

              mean median stddev  CI.1 CI.2
combined.cd 1.47   1.47   1.41 0.765 2.97
```

This gives an OR of 1.47 with 95% CI of (0.765, 2.97), which again indicates that there is no statistically significant difference between the rosiglitazone group and the control group on cardiovascular death.

8.4 Discussion

In this chapter, we discussed meta-analysis of rare events based upon the well-known rosiglitazone dataset using the novel *confidence distribution* approach developed to unify the framework of meta-analysis. We pointed out that the classical fixed-effects and random-effects models are not appropriate for rare events. We recommend the new *confidence distribution* procedure which can combine test results based on exact distributions. The application of this new procedure is made easy with the R package `gmeta`.

For further reading, we recommend Sutton et al. (2002) which provides a review of meta-analyses for rare and adverse event data from the aspects of model choice, continuity corrections, exact statistics, Bayesian methods and sensitivity analysis. There are other newly developed methods for meta-analysis of rare-events. Cai et al. (2010) proposed some approaches based on Poisson random-effects models for statistical inference about the relative risk between two treatment groups. To develop fixed-effects and random-effects moment-based meta-analytic methods to analyze binary adverse-event data, Bhaumik et al. (2012) derived three new methods which include a simple (unweighted) average treatment effect estimator, a new heterogeneity estimator, and a parametric bootstrapping test for heterogeneity. Readers may explore these methods for other applications.

References

Bhaumik, D. K., Amatya, A., Normand, S. T., Greenhouse, J., Kaizar, E., Neelon, B., et al. (2012). Meta-analysis of rare binary adverse event data. *Journal of the American Statistical Association*, *107*(498), 555–567.

Cai, T., Parast, L., & Ryan, L. (2010). Meta-analysis for rare events. *Statistics in Medicine*, *29*(20), 2078–2089.

Chen, D. G., & Peace, K. E. (2013). *Applied meta-analysis using R*. Chapman & Hall/CRC.

Dahabreh, I. J. (2008). Meta-analysis of rare events: an update and sensitivity analysis of cardiovascular events in randomized trials of rosiglitazone. *Clinical Trials*, *5*, 116–120.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, *17*, 841–856.

Harwell, M. (1997). An empirical study of hedgeï£¡s homogeneity test. *Psychological Methods*, *2*, 219–231.

Lane, P. W. (2012). Meta-analysis of incidence of rare events. *Statistical Methods in Medical Research* (2012 Jan 4 Epub).

Nissen, S. E., & Wolski, K. (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *The New England Journal of Medicine*, *356*, 2457–2471.

Raudenbush, S. W. and A. S. Bryk (1985). Empirical bayes meta-analysis. *Journal of Educational Statistics*, *10*(2), 75–98.

Shuster, J. J., Jones, L. S., & Salmon, D. A. (2007). Fixed vs random effects meta-analysis in rare event studies: the rosiglitazone link with myocardial infarction and cardiac death. *Statistics in Medicine*, *26*(24), 4375–4385.

Shuster, J. J., & Schatz, D. A. (2008). The rosigliazone meta-analysis: Lessons for the future. *Diabetes Care*, *31*(3) (10 March 2008).

Sutton, A. J., Cooper, N. J., Lambert, P. C., Jones, D. R., Abrams, K. R., & Sweeting, M. J. (2002). Meta-analysis of rare and adverse event data. *Expert Review of Pharmacoeconomics and Outcomes Research*, *2*(4), 367–379.

Sweeting, M. J., Sutton, A. J., & Lambert, P. C. (2004). What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, *23*, 1351–1375.

Tian, L., Cai, T., Pfeffer, M., Piankov, N., Cremieux, P., & Wei, L. (2009). Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 by 2 tables with all available data but without artificial continuity correction. *Biostatistics*, *10*(2), 275–281.

Chapter 9
Missing Data

Steven A. Gilbert and Jared C. Christensen

> Investigators, sponsors, and regulators should design clinical trials consistent with the goal of maximizing the number of participants who are maintained on the protocol-specified intervention until the outcome data are collected (National Research Council 2010).

9.1 Introduction

Missing data in clinical trials are defined as planned information that was not collected. Examples of these include a subject withdrawing consent before the end of a trial or a laboratory test that cannot be obtained. Depending on why and how much data are missing, the results and interpretability of the trial can be jeopardized. Fortunately, there is a vast literature about statistical methods that can handle missing data. Probably the most important—and least technical—point in this literature is to do everything possible beginning at the trial design stage to avoid missing data. Even when the trial is designed to minimize missing data, the analysis plan should address how the analysis will proceed in the presence of missing data. This chapter should be viewed as a springboard into understanding the minimal amount of theory needed to apply these methods and begin reading the source literature.

In order to have a short exposition of the necessary theory, we will limit most of or discussion to analyses with missing continuous response data. Missing covariates will not be addressed since covariates in clinical trials are almost always baseline variables that are either completely collected, or not collected at all. Though missing

S. A. Gilbert (✉)
Early Clinical Development, Pfizer Inc., Cambridge, MA, USA
e-mail: Steven.A.Gilbert@Pfizer.com

J. C. Christensen
e-mail: Jared.Christensen@Pfizer.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_9

covariates may be an important factor in epidemiological and other studies, they are not addressed in this chapter.

While this chapter is a review of common methods; its unique aspect is in organizing the material around two common overlapping themes:

1. Missing data can require the consideration of multiple models that we call the **analysis** model, **missingness** model and the **complete data** model.
2. Missing data is best viewed in terms of an entire data set and thought of as incomplete data. This allows us to 'borrow' information from observed variables to reduce bias and increase efficiency. Mathematically, this often results in writing down a likelihood for the complete data, that is both observed and missing, and then integrating (i.e., averaging) out the missing data. Therefore, instead of only considering missing responses and observed responses we also consider covariates and auxiliary variables. Auxiliary variables help predict who will have missing data or predict which observations will be missing but are not included in the analysis model.

In the sections that follow we link these themes to methods from the literature. A running example is used to show that the choices made for the three models and how they are used to obtain a treatment effect (Sect. 9.8) will effect the final results and requires thoughtful consideration.

9.2 Preliminaries

Before explaining the multiple models enumerated above, we introduce some basic concepts and nomenclature.

9.2.1 Monotone Versus Non-monotone Missing Data

The longer a trial runs, the more opportunity there is to have missing data. Many trials measure an outcome at multiple times producing longitudinal or repeated measures data. Each subject contributes multiple observations producing a trajectory over time. Even a study with only baseline and follow-up can be considered a repeated measures study.

If measurements are taken as planned and then stop completely before the study ends, the subject is said to have a monotone missing data pattern. When measurements are missing at an intermediate time point followed by observed measurements at least one additional time point, the missing data pattern is said to be non-monotone. As an example consider a publicly available synthetic data set based on a clinical trial for the treatment of depression available at www.missingdata.org.uk. In this trial subjects had their Hamilton Depression Rating Scale (HAM-D) measured at baseline and

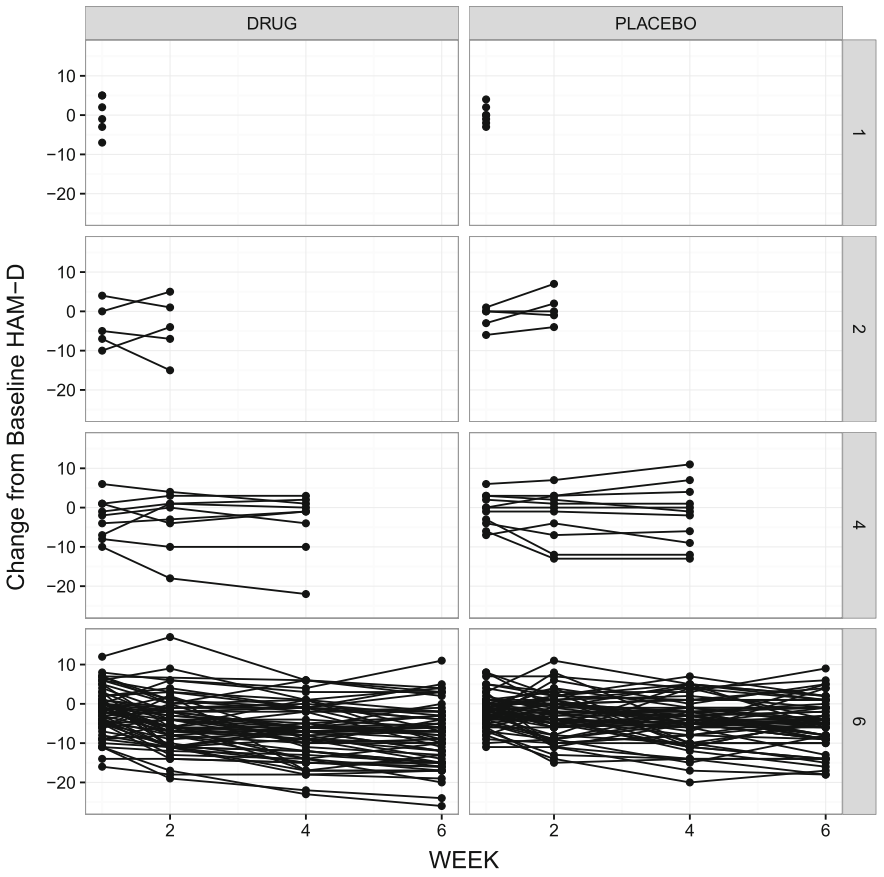


Fig. 9.1 Monotone missing data

weeks 1, 2, 4 and 6. As seen in the spaghetti plot, Fig. 9.1, with separate panels for each combination of treatment (Drug or Placebo) and last visit (week 1, 2, 4 or 6), the data set has all monotone missing data. The panels on the bottom row are the completers, the panels above that group subjects by the time of their last visit.

The distinction between monotone and non-monotone is important for three reasons. First, not all missing data methods can be applied to non-monotone missing data. Secondly, non-monotone missing data is often handled in a two part fashion, first dealing with the early gaps in the data and then going on to use methods that work with monotone missing data. Third, monotone and non-monotone missing data patterns can be created by different underlying causes. Subjects may intermittently miss visits to a clinic for reasons unrelated to their health status, for example, inclement weather; while withdrawing from a trial early is more likely related to their health and study drug (e.g., an adverse event or an exacerbation of their symptoms).

9.2.2 Missing Data Versus Missing Information

Missing data does not necessarily imply we are missing information, particularly in 'outcome' trials. Trials can be grouped into outcome trials and symptomatic trials (O'Neill and Temple 2012). Outcome trials evaluate easily measured endpoints such as mortality in cardiac subjects. While our depression trial or a trial for a skin disease, plaque psoriasis, would be classified as symptomatic trials.

As a hypothetical example, suppose subjects with a history of heart disease were enrolled in a trial testing a new cardiac drug. The primary endpoint is a 6 minute walk test, where a subject is asked to walk quickly on a hard surface for 6 minute and the total distance walked is recorded. Subject A withdraws consent prior to the last visit and leaves the trial. Subject B dies from a cardiac arrest before their last visit. Subject A has missing data and provides no information for their final visit, subject B on the other hand provides compelling information about their cardiac status at their last visit even though they have missing data.

In contrast, consider the same scenario in a plaque psoriasis trial measuring the Psoriasis Area Severity Index (PASI), a measure of the extent and severity of a subject's skin lesions. Again subject A withdraws consent and subject B dies from a cardiac arrest. Now neither subject provides data or information on the course of their plaque psoriasis. Scientific judgment is necessary to put the data into context, statistical methods alone are not sufficient.

9.2.3 Notation

We now consider notation that can describe longitudinal data and distinguish between a vector of values, denoted with capital letters and single data points denoted with lowercase letters. This notation can also be used for single responses as well if vectors of length 1 are used. The complete response data is $Y = \{Y_1, \ldots, Y_n\}$ where each Y_i is a vector containing the complete data on the endpoint of interest for subject i and y_{it} is the response for subject i at time t. Similarly $R = \{R_1, \ldots, R_n\}$ contains the response indicator for all subjects at all time points with r_{it} the response indicator equal to 1 if the data are observed for subject i at time t and 0 otherwise.[1]

It is convenient to have notation separating the observed from the missing data; define $Y = \{Y_O, Y_M\}$ where $Y_O = \{Y_{1O}, \ldots, Y_{nO}\}$ are the observed responses for subjects $i = 1, \ldots, n$ and $Y_M = \{Y_{1M}, \ldots, Y_{nM}\}$ are the unobserved or missing responses for these subjects. We will refer to $Y_i = \{Y_{iO}, Y_{iM}\}$ as the complete data for subject i. Furthermore, let $X = \{X_1, \ldots, X_n\}$ contain the treatment assignment and other baseline covariates.

[1]The model can be extended further by allowing R to take on more than two values indicating multiple response patterns. This will be necessary for applying pattern mixture models.

9.3 Analysis Model

Using the depression data as an example begin by examining basic summary statistics of change from baseline HAM-D and baseline HAM-D grouped by gender and drug treatment; these summaries are displayed in Table 9.1. First note that about 25% of the total response data are missing. This is a substantial amount and cannot be ignored. Furthermore, the male placebo subjects have the highest rate of missing data ($100\% - 69\% = 31\%$) and the smallest decrease from baseline, -4.73. They also have the lowest mean baseline depression score, 16.97.

Based on the summary statistics in Table 9.1 a treatment effect contrast can be calculated. Using equal weights the contrast is:

$$(0.5 \times -5.35 + 0.5 \times -4.73) - (0.5 \times -8.89 + 0.5 \times -7.69) = 3.25.$$

Contrast weights calculated with randomized proportions (47/84, 37/84, 56/88, 32/88) or observed proportions (35/64, 29/64, 43/65, 22/65) provide similar results.

Table 9.1 also shows there are small differences in the average baseline HAM-D scores in gender by treatment groups. Examining the relationship between baseline and change scores can be done with Fig. 9.2, where there is a relationship between baseline and change scores only in the treated subjects.

Because change from baseline HAM-D at week 6 can be regarded as a continuous variable we choose to analyze the data with an analysis of covariance (ANCOVA), with treatment, baseline HAM-D and gender as fixed effects. Using **R**'s modeling syntax, the model can be written as,

$$Change \sim Drug + Gender + Baseline,$$

indicating that the categorical variables, Drug (Drug or Placebo), Gender (Male or Female) and the continuous baseline HAM-D score are all main effects.

A naive approach is to ignore the missing data and analyze only the complete cases (CC), that is we fit a linear model to the subjects that have data through week 6 as follows:

Table 9.1 Summary of change from baseline scores

Drug	Gender	N	Nobs	Mean baseline	Mean change	SD change	Percent observed (%)
DRUG	F	47	35	18.77	−8.89	8.42	74
DRUG	M	37	29	18.46	−7.69	6.10	78
PLACEBO	F	56	43	17.32	−5.35	6.31	77
PLACEBO	M	32	22	16.97	−4.73	5.91	69

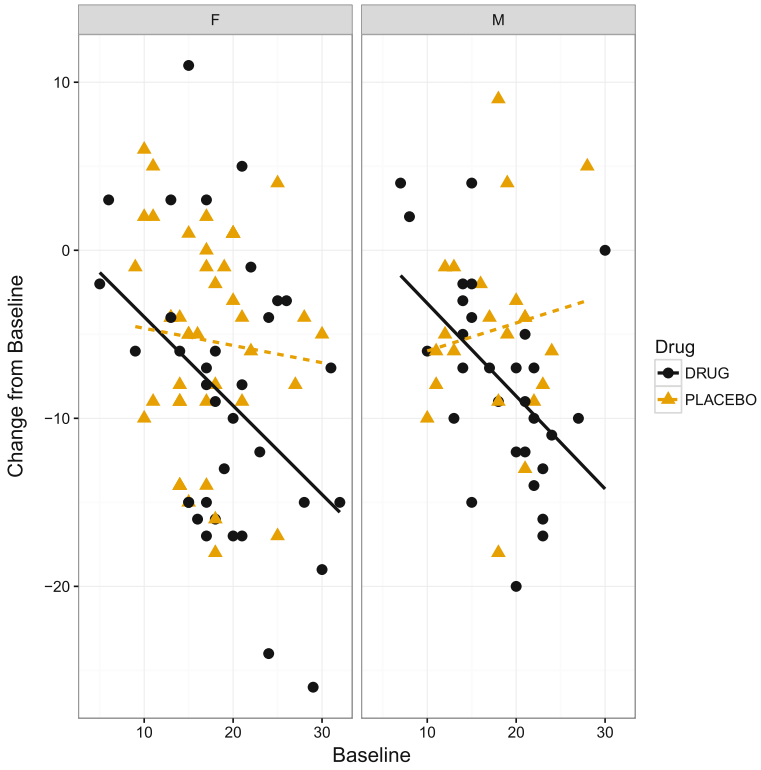


Fig. 9.2 Depression data by gender and treatment

```
lmcc <- lm(Change_wk6 ~ Drug + Gender + Baseline)
```

The estimated treatment difference from the model is taken as the β coefficient for the drug effect, an overall contrast comparing drug to placebo. The resulting treatment contrast from this analysis is displayed in Table 9.2 along with treatment contrasts from three reduced models. The model with only a drug effect which directly compares the average change scores without adjustments for baseline or gender is 3.21, almost identical to the contrast derived above from summary Table 9.1.

Table 9.2 Summary Complete cases treatment differences

Model	Estimate	Standard error
Drug	3.21	1.20
Drug + Baseline	2.66	1.17
Drug + Gender	3.31	1.21
Drug + Gender + Baseline	2.76	1.19

Including gender in the model increases the contrast slightly, however, the two models that adjust for baseline have significantly lower treatment contrasts, 2.66 and 2.76. We return to the reason for these differences later, after introducing the missingness and complete data models.

9.4 A Model for Missingness

A model for missingness is simply a model for the responder variable R. In practice this will be a logistic regression or similar model such as probit regression. As an example consider the following logistic model with a three way interaction between the continuous baseline score, drug and gender with an additional additive term for site included to illustrate the fact that the missingness model can contain different covariates than the analysis model. The model can be fit with the following **R** code,

```
mmodel1 <- glm(r ~ Drug:Gender:Baseline + Site, family=binomial(link="logit")).
```

The first use we have for this model is inverse probability weighting, the topic of the next session.

9.5 Inverse Probability Weighting

Inverse probability weighting (IPW) was proposed for survey inference by Horvitz and Thompson in the 1950s and is also used in Monte Carlo simulations under the guise of importance sampling (Kang and Schafer 2007).

To gain intuition, suppose we can obtain a random sample, $\{y_i, \ldots, y_n\}$ drawn from a continuous distribution $f(y)$, then the mean of the distribution can be estimated using the sample mean,

$$\mu_f(y) = E[Y] = \int y f(y) dy \simeq \frac{1}{n} \sum y_i.$$

Now suppose we want the mean under a different distribution, g, then consider the following,

$$\begin{aligned} \mu_g(y) &= \int y g(y) dy \\ &= \int y \frac{g(y)}{f(y)} f(y) dx \\ &\simeq \frac{1}{n} \sum \frac{g(y_i)}{f(y_i)} y_i \\ &= \frac{1}{n} \sum w_i y_i, \end{aligned}$$

where $w_i = g(y_i)/f(y_i)$. Each w_i is a ratio of the probabilities of y_i being drawn from density g instead of f. Values that are likely under g get large weights, while unlikely values get small weights in just the right proportion so that expected values are consistent estimates of their population average under g. Therefore, the weighting can be used to take samples drawn from one distribution and calculate a mean or any expectation under a second distribution, if the distributions have the same support. When data are missing, this method can be used to take the complete cases who may no longer represent the entire population and adjust the mean to represent the population.

In practice, a logistic regression or similar model is fitted to the response indicator to obtain predicted probabilities π_i that each observation will be observed conditional on the terms in the logistic regression model. Observations with a small π_i are very valuable (they are not observed often), therefore they require a large weight calculated as $w_i = 1/\pi_i$. There are two basic estimators that incorporate these weights;

$$\hat{\mu}(y)_{IPW1} = \frac{1}{n}\sum \frac{r_i y_i}{\pi_i} = \frac{1}{n}\sum w_i r_i y_i, \tag{9.1}$$

$$\hat{\mu}(y)_{IPW2} = \left[\sum \frac{r_i}{\pi_i}\right]^{-1} \sum \frac{r_i y_i}{\pi_i} = \left[\sum w_i r_i\right]^{-1} \sum w_i r_i y_i, \tag{9.2}$$

which are both consistent for μ if the weights are correctly modeled (Cao et al. 2009)

Returning to the depression example, we use logistic regression to model the probability that each patient has week 6 data, $P(r_i = 1)$. The probabilities are output from the model and transformed into inverse probability weights $w_i = 1/P(r_i = 1)$,

```
weight1 <- 1/predict(mmodel1, type="response")
```

which can then be used as weights to a regression function.

It is good practice to examine the weights used in an analysis. Boxplots of the weights for the depression analysis are seen in Fig. 9.3, along with the average weight displayed in each box. It is clear that the male placebo group which had the greatest missing data rate has the highest weights on average. The intuitive justification is that the weights are recreating the original sample, assuming that the missing data looks like the observed data. For example, if we only used the treatment and gender information in the model, then the probability of observing data for any of the male placebo subject is $22/32 \approx 0.69$, so that the weights would be $32/22 \approx 1.45$ for each male placebo subject. That is each completely observed male placebo subject counts for 1.45 subjects in the analysis.

Looking again at Fig. 9.3 one possible drawback of this method is some outlying weights. An IPW analysis can suffer from the undue influence of a small number of large weights. In addition, the further the weights differ from equal weighting, $1/n$, the greater the standard error of the mean estimate.

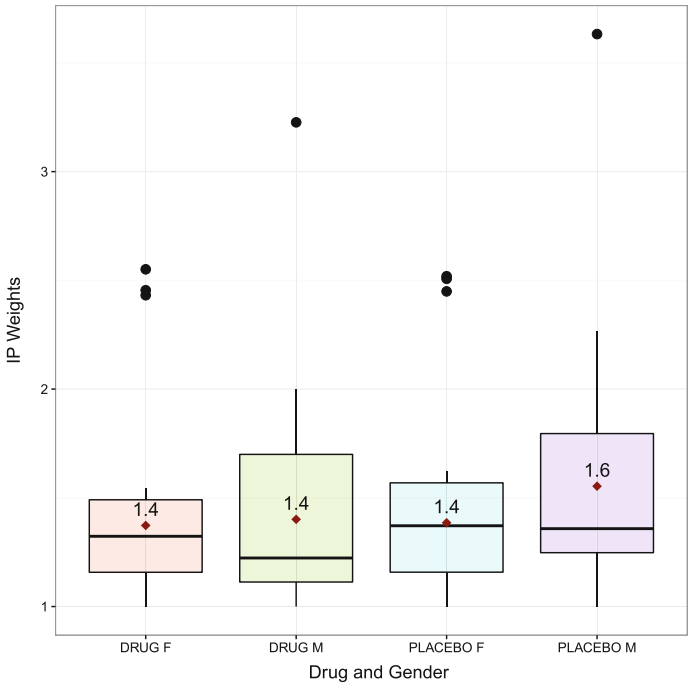


Fig. 9.3 Inverse probability weights

As described earlier an IPW analysis can be combined with the analysis model by introducing the weights into a standard regression. For example, the weights can be used with an analysis of variance model (ANOVA) as follows,

```
ipw1 <- lm( Change_wk6 ~ Drug, weights= weight1) .
```

The regression output has a mean difference of 3.3 and standard error of 1.21. However, the standard error is not correct because of the use of estimated weights. The regression analysis treats the weights as fixed known quantities and does not adjust for the fact they are estimated and not known. Analytical calculations of the standard error are difficult but numerical estimation with a bootstrap algorithm can be implemented. Using a nonparametric bootstrap, the estimated standard error is slightly higher, 1.31, as expected.

The IPW weighting can also be applied to models with covariates. Using these weights with the original analysis model including Drug, Gender and Baseline, resulted in a estimated mean difference of 2.9 and a bootstrapped standard error of 1.34.

Like the complete case estimates using the analysis model alone in the previous section, the results change depending on our choices for the missingness and analysis models. We return to this topic later, but first look at complete data models.

9.6 Complete Data Model

A model for the complete data describes both the observed and missing data. This may or may not be the same as the analysis model. An analysis model could be much simpler, such as an ANOVA model with only a treatment effect, ignoring all other variables. The complete data model makes use of the second theme from the introduction, viewing missing data as incomplete data. The entire data set has missing values, but the data has structure and that structure allows us to borrow information from observed data to help infer what happens in unobserved data.

As an illustration for the depression data consider a linear model that allows a different slope for each combination of gender and drug treatment. This model can be fitted in **R** as follows,

```
cd1 <- lm( Change_wk6 ~  Drug:Gender:Baseline),
```

Predicted values can be calculated for all subjects based on this model and used to calculate a treatment contrast,

```
Predicted <-   predict(cd1, hamd.la)
reg1 <- lm(Predicted ~ Drug)
```

The treatment contrast is calculated to be 3.6. The reported standard error from the regression on the predicted values does not accurately reflect the true standard error, since information on the variability of the data is lost when looking only at fitted values which do not include random error. As before, a bootstrap can used to obtain a valid estimate of 1.6.

The analysis was repeated using a complete data model with a single slope,

```
cd2 <- lm( Change_wk6 ~  Drug + Gender + Baseline) .
```

The treatment contrast is now 3.16 with a standard error of of 1.17.

9.7 It Is All About the Weighting

The missing data methods examined so far all apply different weightings to the observed responses. The weighting methods obviously specify weights for each observation in an explicit manner. Linear models do so in an implicit manner. Recall that for a linear model,

$$Y = X\beta + e,$$

where; Y is an $n \times 1$ column vector of responses, X is an $n \times p$ design matrix, e is an $n \times 1$ column vector of error terms with $E(e) = 0$ and $Var(e) = V$ an $n \times n$

covariance matrix. The solution of the normal equation for the parameter vector β is,

$$\hat{\beta} = (X^T V^{-1} X) X^T V^{-1} Y. \tag{9.3}$$

The $p \times n$ matrix

$$(X^T V^{-1} X) X^T V^{-1},$$

has a row for each β parameter and a column for each observation y_i, therefore, the regression estimates (i.e., the β's) are calculated as weighted sums of the observations. If V is a diagonal matrix and all error terms have equal variability, (i.e., $V = \sigma^2 I$), where I is an $n \times n$ identity matrix, the V^{-1} terms cancel out and

$$\hat{\beta} = (X^T X) X^T Y.$$

The β estimates are still weighted sums of observations, but now the weights are determined by the X matrix alone.

Suppose we want to estimate a mean response of a simple random sample, $Y = \{y_1, \ldots, y_n\}$ using IPW weights. The linear model framework above can be used with a single vector of 1's as the design matrix and replacing V^{-1} with a diagonal matrix of weights, W, in Eq. 9.3 to obtain

$$\hat{\beta} = (X^T W X) X^T W Y.$$

showing that adding weights is equivalent to specifying a particular form of heteroscedasticity. When considering longitudinal data, correlations between observations on a single subject need to be accounted for, therefore the calculations will need to incorporate both a correlation structure and weights. This can done by replacing V in Eq. 9.3 with $W^{-1/2} V W^{-1/2}$.

Weighting of the observations is even more transparent when examining the fitted or predicted values used in Sect. 9.6,

$$\hat{Y} = (X^T V^{-1} X) X^T V^{-1} Y = HY,$$

where H is the $n \times n$ hat matrix that transforms the data Y to estimates $\hat{Y}$ by matrix multiplication. Each row of H is an n vector $\{H_{i1} \ldots, H_{in}\}$, where H_{ij} is the row i, column j element of H. The entries of H define each $\hat{y}_i$ as weighted average of the observed y_i as follows,

$$\hat{y}_i = \sum_{j=1}^{n} H_{ij} y_j.$$

There is a further level of weighting in the Sect. 9.6 analysis induced by averaging over predictions for all subjects including those with missing data. This re-weights the contrast to reflect the original randomized proportion in each treatment arm instead of the observed proportions in the CC data.

The point of this discussion is that regression modeling and IPW weighting both produce estimates by calculating weighted linear combinations of the data. Regression methods highlight the contributions of the variables in the analysis and define the weights implicitly while weighting methods focus on the weights and use covariates implicitly. However, both methods depend heavily on our choice of models and how we use the models to calculate a treatment contrast, the topic of the next section.

9.8 Treatment Contrast

Let us return to the question of why the treatment contrasts illustrated so far differ from one another. Empirically, the treatment differences depended primarily on whether baseline was included in the modeling, while gender had a minor effect. A graphical representation of two models for change from baseline that depend only on drug and baseline value can shed some light on what is happening.

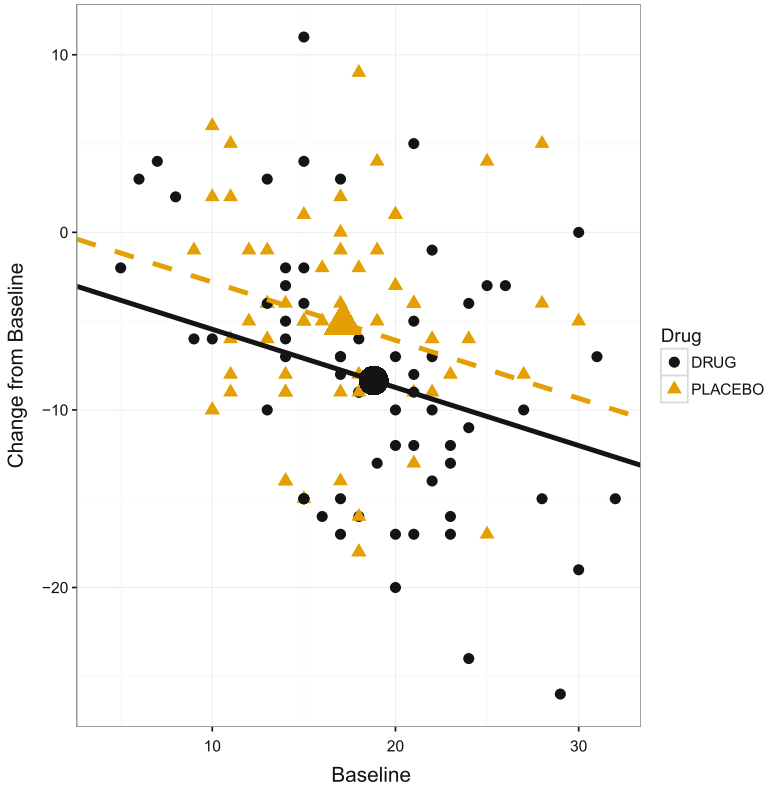


Fig. 9.4 ANCOVA with single slope

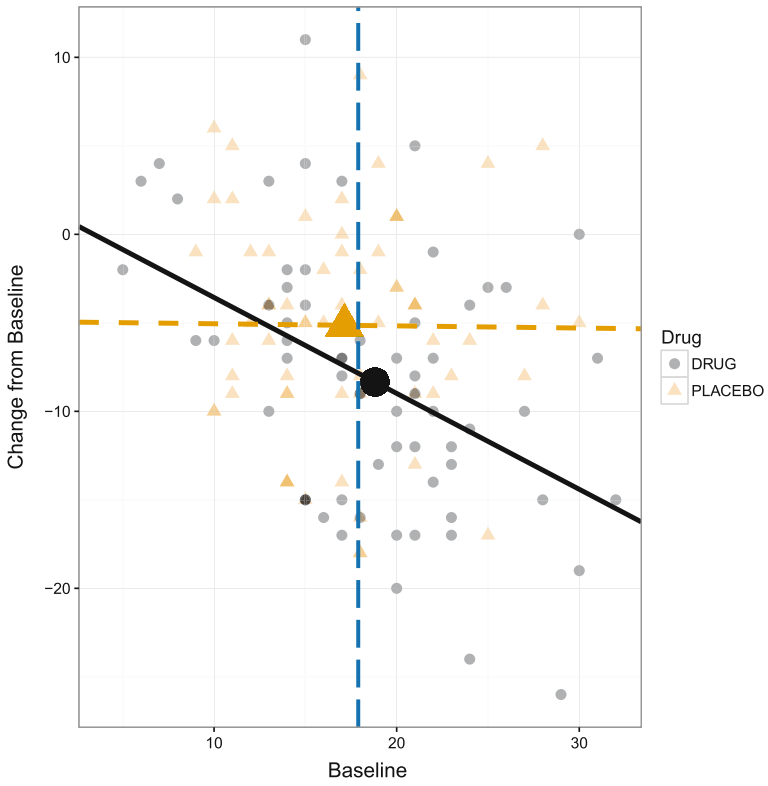


Fig. 9.5 ANCOVA with two slopes

The first model, graphed in Fig. 9.4, constrains both regression lines to a single slope, while the second model, graphed in Fig. 9.5, allows a different slope for each drug group (note that the placebo slope is very close to zero in the figure). The figures have the individual data as well as the drug group specific regression lines. The larger filled in circle and triangle lying on the regression lines are placed at the average baseline score for each group indicating the predicted change score averaged over the baseline values for a treatment group.

For linear models and generalized linear models (GLMs) with an identity link, the average of the response is equal to the response calculated at the average value of the covariates. For example, consider the average of the placebo subjects predicted values,

$$\begin{aligned}\frac{1}{n_{Plb}}\sum_{i\in Plb}\hat{y}_i &= \frac{1}{n_{Plb}}\sum_{i\in Plb}\beta_{Int}+\beta_{Plb}+\beta_{Slope_{Plb}}baseline_i\\ &= \beta_{Int}+\beta_{Plb}+\beta_{Slope_{Plb}}\frac{1}{n_{Plb}}\sum_{i\in Plb}baseline_i\\ &= \beta_{Int}+\beta_{Plb}+\beta_{Slope_{Plb}}\bar{\mu}_{Baseline_{Plb}}.\end{aligned}$$

Making the same calculation for treated subject would result in

$$\beta_{Int}+\beta_{Drug}+\beta_{Slope_{Drug}}\bar{\mu}_{Baseline_{Drug}}.$$

It can be argued that a fair treatment comparison in a controlled clinical trial should hold the baseline score fixed, since we want to isolate the effect of treatment from all other factors. In terms of the ANCOVA models, this means looking at the vertical separation between the regression lines at a fixed baseline value on the x-axis. If there is a single slope, as in Fig. 9.4, the result is the same for any baseline value, as long as it is the same for both treatment groups. This type of comparison is done automatically when comparing groups with least squares means in **SAS**. Even with a single slope, if the treatment contrast is made using different baseline averages the treatment difference can be made larger or smaller depending upon where the averages fall. Figure 9.4 shows that using the observed treatment specific baseline averages would increase the treatment contrast.

Different slopes, as in Fig. 9.5, present a more difficult problem. Even if the baseline score is held fixed across treatment groups, the treatment difference will vary widely depending upon where the comparison is made. Applying the least squares means strategy, contrasting the groups at the average baseline HAM-D score of 17.90 (dashed vertical line) in all subjects, the estimated treatment contrast is 2.70.

To sum up, even after an analysis model and missing data strategy have been chosen, the details of how the analysis model will be used to estimate the treatment contrast still needs to be carefully chosen.

9.9 Selection Model Factorization

A full likelihood based approach contains a probability model for both the complete data and the missingness process that are combined into a single joint likelihood. A selection model factorization of this joint likelihood is used to classify missing data as missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). In practice, MCAR, MAR and MNAR determine how difficult it is to analyze the data with missing values and what type of statistical methods are appropriate. If the data are MCAR, the missing data can be ignored and any type of modeling, including summary statistics such as the mean and median provide consistent estimates of the population parameters. If the data are MNAR, the observed data alone is not enough to provide consistent estimates of population parameters; un-testable assumptions about the missing data are needed to obtain estimates. MAR

is a middle ground where simple summary statistics are not consistent but correctly specified likelihood based analyses can be used ignoring the missing data mechanism.

The selection model factors the joint likelihood of the complete data and response indicators as follows:

$$P[Y_O, Y_M, R|X] = P[R|Y_O, Y_M, X] \times P[Y_O, Y_M|X]. \tag{9.4}$$

The first term on the right hand side (RHS) determines the probability that an observation is observed (i.e., the missingness model) while the second term on the RHS is the complete data model.

The distinctions between MCAR, MAR and MNAR are clearer if you consider using (9.4) to build a Monte Carlo simulation of missing data. Start with a fixed design matrix of covariates X; next generate the complete data, $Y = \{Y_O, Y_M\}$ from the complete data model $P[Y|X]$. The last step is to use the missingness model as a filter to select which of simulated y_i's will be retained in the data set. This is where MCAR, MAR and MNAR come into play.

1. The model is MCAR if the missingness model is does not depend on X or Y. Equivalently, the observed data are a random sample from the complete data. Any type of analysis can be used to obtain an unbiased estimate, however, efficiency is decreased because the sample size has been reduced.
2. The model is MAR if the missingness model depends on X and Y_O but not Y_M or other unobserved variables. If in addition, the parameters of the complete data and missingness models are distinct, the missing data mechanism is ignorable and the complete data model can be fitted alone without incurring bias.
3. The model is MNAR if the missingness model depends on Y_M. This is the most difficult case since the missingness model depends on unobserved data. Inference can only proceed under strong assumptions that cannot be tested.

The results enumerated above are obtained by marginalizing the likelihood to remove the contribution of the missing data. Marginalizing is done by integrating out Y_M which in general can be difficult but, under the selection model assuming the data are MCAR or MAR, simplifications occur.

$$\begin{aligned} P[Y_O, R|X, \theta] &= \int P[Y_O, Y_M, R|X, \theta] dY_M \\ &\overset{MAR}{=} \int P[Y_O, Y_M|X, \theta] P[R|Y_O, \eta] dY_M, \\ &\overset{MCAR}{=} \int P[Y_O, Y_M|X, \theta] P[R|\eta] dY_M, \end{aligned}$$

where we now include the parameters θ and η of the complete data and missingness models. Under MAR and MCAR, the missingness model does not depend on Y_M and can pass through the integral while Y_M can be integrated out of the complete likelihood. We show this for MAR,

$$\begin{aligned}\int P[Y_O, Y_M|X, \theta]P[R|Y_O, Y_M, \eta]dY_M &= \int P[Y_O, Y_M|X, \theta]P[R|Y_O, \eta]dY_M \\ &= P[R|Y_O, \eta]\int P[Y_O, Y_M|X, \theta]dY_M \\ &= P[R|Y_O, \eta]P[Y_O|X, \theta].\end{aligned}$$

The likelihood now has two terms, if in addition, θ and η are distinct, the likelihood of the complete data model can be maximized independently from the missingness model. Since in most cases the missingness model is not of interest, the complete data model can be fitted as usual ignoring the missing data model completely.

Selection models are not often fitted directly for clinical trials data. However, an argument that the data are MAR and the parameters are functionally distinct, or equivalently the missing data are *ignorable* can be used to justify modeling methods such as a mixed model repeated measures analysis (MMRM) which we take up in the next section.

9.10 Linear Mixed Model Repeated Measures Analysis

Mixed model repeated measures (MMRM) analyses have been recommended when the missing data are MAR (National Research Council 2010). These models define a joint probability distribution over the complete response vector and naturally incorporate the incomplete data view, borrowing information from observed responses to help make inference where data are missing. Studies have shown that MMRM analyses can work well in the presence of missing data (Mallinckrodt et al. 2001).

The use of an MMRM takes advantage of the fact that the MMRM can function as both the complete data and analysis model. Although, these are regression models they define a multivariate distribution for the responses over time, conditional on the covariates. The assumption that the response at an earlier post-baseline visit is informative about a missing visit later in the trial is reasonable. An important point is that the model includes information from all observed visits but does not condition on that data. To be more explicit, week 4 change scores are likely to be predictive of week 6 change scores in the depression data. Week 4 scores can be included in the MMRM as a response, Y variable, without changing the interpretation of a treatment contrast. On the other hand if week 4 scores were added to an ANCOVA for change at week 6 as a covariate, they would change the interpretation of the treatment comparison. The MMRM can work with monotone and non-monotone missing data and fits the data in a single step (i.e., a separate missingness and imputations models are not needed).

There are some shortcomings to the MMRM approach. The first is that we still need to minimize the number of covariates to avoid complicating the interpretation of the treatment contrast. Even more important is that post-baseline data cannot be included as covariates. Incorporating as much auxiliary data as possible is referred

to as 'inclusive' modeling as opposed to 'restrictive' modeling (Collins et al. 2001). MMRM analyses encourage a restrictive approach while multiple imputation (MI), to be covered shortly, encourages a more inclusive approach.

Next, we define a linear MMRM and present results for the depression data. A linear MMRM model can be defined as

$$\begin{aligned} Y &= X\beta + Z\gamma + e, \\ e &\sim N(0, P) \\ \gamma &\sim N(0, G) \end{aligned}$$

where we are again using vector and matrix notation. The γ term is vector of random effects,while Z is a design matrix for random effect. The covariance matrix, V for Y now has the structure,

$$V = Var(Y) = Z^T GZ + P.$$

If the model contains Z and γ but P is a diagonal covariance matrix[2] it is called a random effects model. Random effects for categorical variables such as subject can be coded in a Z matrix with entries entirely made of 0's and 1's and are called variance components. These components model clustering of data well, where groups of observations share a positive correlation. However, they do not model correlations that change over time. For example you would expect the responses at weeks 2 and 4 to be more highly correlated than weeks 2 and 6. These types of correlations can be modeled using correlation structures from time series analysis such as an autoregressive (AR) structure. These time series structures describe correlations within a single subject and are therefore included in the P matrix in a block diagonal fashion. The block diagonal structure models correlations between visits for a single subject but keeps data from different subjects independent.

Software to fit MMRMs is readily available. The following code uses the nlme package in **R**

```
mixedfit3 <- lme(change ~ WEEK +Drug:WEEK +
    Baseline:WEEK,random= ~1|SUBJID,
cor=corSymm( form=~1|SUBJID) ,
control=list(opt="optim"))
```

The MMRM above (where we have dropped gender to simplify the model) allows for an arbitrary correlation structure over time in addition to a random subject effect. We choose a flexible correlation structure because that is the part of the model that determines how information will be borrowed from the observed data. The resulting treatment contrast is 2.72 with a standard error of 0.95. This is all reported directly from the model fit, bootstrapping or other post-fitting procedures are not needed.

[2]Common notation for the $Var(e)$ is R but has been changed to P to avoid confusion with the response vector R.

The MMRM approach requires the missing data to be MAR. This translates into the assumption that subjects with missing week 6 data, have unobserved responses similar to other subjects in their treatment group with similar covariates and change from baseline trajectories up until the time of dropout. Clinically, this assumes that subjects with missing data have similar responses to subjects who stay on protocol and maintain their drug effect. This may not be a good assumption if subjects cannot tolerate a drug intended for a chronic condition since they will receive no benefit from the drug after withdrawal.

9.11 Generalized Linear Mixed Models

Binary data is often analyzed in clinical trials. For example did the subject have a 30% decrease from baseline, yes or no. It is tempting to claim the data are MAR and use a logistic version of the MMRM. Unfortunately, these models estimate a different effect than linear mixed models. A little background information is needed to understand why this is so.

Linear mixed models specify a model for the observations directly, while GLMMs specify a model for the mean structure directly,

$$E[Y|\gamma] = g^{-1}(X\beta + Z\gamma), \tag{9.5}$$

where $g(\cdot)$ is the link function, a monotone invertible function linking the mean to a linear function of the predictors, $X\beta + Z\gamma$, called the linear predictor. The crux of the issue is that

$$E[g^{-1}(X\beta + Z\gamma)] \neq g^{-1}(X\beta + ZE[\gamma]), \tag{9.6}$$

unless $g(\cdot)$ is the identity function, then

$$E[X\beta + Z\gamma] = X\beta + ZE[\gamma]. \tag{9.7}$$

The implication of this, is that for linear mixed models the β coefficients can be interpreted at both a subject level and a population level. Consider a model with an additive treatment effect, β_{Drug} and random subject effect. Conditional on the random subject effect, the treatment effect is β_{Drug}; it is also the population effect (how large is the treatment over all subjects) since the random subject effect can be averaged out without effecting β_{Drug}. This is not so for other link functions. The GLMM specifies the β coefficients as the effect conditional on the other predictors in the model including random effects. To get a population level estimate, the random effect needs to be averaged out. Unfortunately, a non-linear link function will, for a lack of a better term, 'link' the random effect with the treatment effect.

The goal of clinical trials is to estimate population level effects, therefore the use of GLMMs requires care. For example, suppose binary data were analyzed, the approach used in Sect. 9.6 could be taken by calculating the predicted probability

of response for each subjects and averaging results on the probability scale instead of on the linear predictor scale. Unfortunately, this detracts from a major advantage of the MMRM approach which its ease of use. Generalized estimating equations (GEEs) can be used for longitudinal binary data and directly produce population level estimates. However, the estimates are only valid if the data are MCAR, this can be overcome by combining IPW with GEE into a weighted GEE (wGEE) procedure (Molenberghs and Verbeke 2005) or using multiple imputation. We close this section by noting there are two approaches to wGEE, weighting individual observations and weighting subjects, both are available in commercially available software including **SAS** and other products.

9.12 Multiple Imputation

No discussion of missing data is complete unless it includes multiple imputation (MI). MI is a general procedure and is applicable for almost any type of missing data, not just MAR and missing response data. MI is to be distinguished from single imputation methods such as last observation carried forward (LOCF), baseline observation carried forward (BOCF), mean imputation or other methods that fill in the missing observation once and then analyze the data set as if it were complete. These methods are discouraged because they ignore the uncertainty in the imputed data and underestimate the variance (National Research Council 2010).

Retaining our concentration on missing response data, consider a statistic Q of the complete data of interest, (e.g., the sample mean or treatment contrast) and associated variance estimate U which is a single number if Q is a scalar, (e.g., a variance estimate for a single β coefficient), or a a matrix if Q is vector valued (e.g., the covariance matrix of the entire β coefficient vector). The basic result justifying multiple imputation is,

$$P(Q|Y_O) = \int P(Q|Y_O, Y_M)P(Y_M|Y_O)dY_M, \tag{9.8}$$

(Rubin 1996), using the second theme in the introduction, averaging over the missing data. In practice, data simulated from $P(Y_M|Y_O)$ are used to fill in the missing values, Y_M, and the the complete data $Y = \{Y_O, Y_M\}$ are analyzed as usual. This filling in and analyzing are repeated $B \geq 2$ times recording the results $\{\hat{Q}_1, \hat{U}_1, \ldots, \hat{Q}_B, \hat{U}_B\}$. The results from the individual imputations are combined with 'Rubin's rules',

$$Q_{Imp} = \frac{1}{B}\sum \hat{Q}_i \tag{9.9}$$

$$Var[Q]_{Imp} = U_W + \frac{B+1}{B}U_B, \tag{9.10}$$

where U_W is the average within imputation variability and U_B is the between imputation variability; calculated as,

$$U_W = \frac{1}{B}\sum \hat{U}_i$$
$$U_B = \frac{1}{B-1}\sum (\hat{Q}_i - Q_{Imp})(\hat{Q}_i - Q_{Imp})^T$$

(Rubin 1996).

MI was originally derived from a Bayesian perspective, where we are simulating from the posterior predictive distribution of missing data conditional on the observed data. However, MI is often used in a frequentist setting and sampling distributions are substituted for predictive distributions.

A description of the imputation algorithm itself, instead of the mathematical justification will make the method more understandable. Consider a regression setting with a response vector $Y = \{y_1, \ldots, y_n\}$ that has observed and missing values along with a fully observed covariate matrix X with entries x_{ij}. The regression model for each y_i is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i.$$

1. Fit the regression model to the observed y_i's, retaining the fitted parameters $\hat{\beta}$, covariance matrix of the parameters, $Var[\hat{\beta}] = \hat{\sigma^2} X^T X$ and the variance estimate $\hat{\sigma}^2$.
2. Draw a new set of parameters $\beta^\star$ from $N(\hat{\beta}, \sigma^{2\star} X^T X)$ and $\sigma^{2\star}$ from $\hat{\sigma^2}(n_O - k + 1)/g$ where g is a $\chi^2_{n_O-k-1}$ random variable and n_O is the number of observed y_i.
3. Fill in the missing y_i with

$$y_i^\star = \beta_0^\star + \beta_1 x_{i1}^\star + \cdots + \beta_{ip}^\star x_p + \sigma^{2\star} z_i^\star,$$

   where the $z_i^\star$ are independent standard normal variables.
4. Repeat B times and combine the results.

(SAS Institute 2015).

Note that the algorithm accounts for two sources of variability, the variability in the parameter estimates, $\hat{\beta}$ and $\hat{\sigma}^2$ at step 2, and variability in the data themselves in step 3.

MI algorithms such as the one above can be used for data that are MAR, but MI can also be used for data that are MNAR. The central idea, is that if you can estimate or define the posterior predictive distribution, MI can be applied.

We close with a note about the number of imputations needed for an analysis. We have been concerned about Monte Carlo error since this is a simulation method. Classic results based on efficiency (Rubin 1996) and ignoring Monte Carlo error indicate that 3–5 imputations are sufficient. Others have suggested the number of imputations should equal the percentage of missing data (e.g., if 10% of the data is

missing use 10 imputations). The number of imputations needed to address Monte Carlo error is less clear. If the primary result from a trial is close to statistically significant (e.g. $p = 0.05$), then changing the random number seed for the imputation algorithm can swing the inference in one direction or another if there are an insufficient number of imputations. We suggest either specifying a large number of imputations or increasing the number until the results of interest are stable.

9.13 Pattern Mixture Factorization

Pattern mixture models (PMMs) reverse the order of conditioning of the selection model,

$$P[Y_O, Y_M, R|X] = P[R|X] \times P[Y_O, Y_M, |R, X], \tag{9.11}$$

(Carpenter and Kenward 2012). Pattern mixture models often use a categorical r_i instead of a binary response indicator. This allows grouping of subjects based on time of dropout or reason for dropout. The primary assumption is that the missingness patterns, captured by the r_i can be identified easily and the modeling effort goes into $P[Y_O, Y_M, |R, X]$ the distribution of the complete data conditional on the dropout pattern and covariates. For example in the depression data, the natural categories would be how many weeks the subject remains in the study. In other trials the categories could include the reason a subject withdraws from the study, such as lack of efficacy, drug reaction etc.

In its simplest form, a PMM stratifies the data set by dropout patterns, estimates the response data and finally combines the information back together. However, there may be too many patterns to efficiently draw inference for each pattern. If subjects are observed at k occasions, there are 2^k possible patterns based on when data are observed, including reason for dropout increases the number even further. Therefore patterns may need to combined on scientific and clinical grounds to make the modeling process tractable.

The MCAR, MAR, MNAR taxonomy is specific to selection models and does not carry over to PMMs, therefore they need to be described in an alternative manner. One way to describe PMMs is by 'identifying restrictions' (Molenberghs and Verbeke 2005). Assume monotone missing data, as in the depression example. One set of possible restrictions are complete case missing values (CCMV), neighboring case missing values (NCMV) and available case missing values (ACMV). Viewed from our borrowing information standpoint the restrictions indicate where information is borrowed from to make inference. The CCMV restriction implies missing information is only borrowed from completers. The NCMV restriction borrows from the nearest pattern, for example if you have a missing week 2 visit, information from the subjects with week 4 but not week 6 are used. Lastly, ACMV is the available case missing data restriction which implies that information is borrowed from all patterns with information at that time point, for example if you have a missing week 2 visit, information from the subjects with week 4 *and* week 6 are used. In the special

case of monotone missing data ACMV is "the counterpart of MAR in the PMM context" (Molenberghs and Verbeke 2005). Other restrictions such as missing non-future dependent (MNFD) require dropout to depend on current unobserved data and not future unobserved data are also used. We do not discuss these further since they are not often used in the pharmaceutical industry, but rather proceed to using PMMs for MNAR data.

A combination of PMMs and MI has become popular for sensitivity analyses of clinical trials when data are MNAR. Suppose subjects withdraw from the depression trial because they cannot tolerate the drug. Clinically we assume they no longer benefit from the drug after they withdraw; therefore their missing data is not expected to be similar to other subjects in their treatment group with similar covariates and pre-withdrawal change from baseline scores. This is a violation of the MAR assumption, hence the data is MNAR. This scenario is well suited to a 'controlled imputation' approach (Mallinckrodt et al. 2012) where information from a reference group is used to impute the missing values. One such approach, jump to control, imputes missing week 6 change scores from the distribution of week 6 change scores in the placebo group. There are many variations of this idea and they need to be justified on clinical and scientific grounds. Once data are declared MNAR, the analysis becomes highly dependent on the models used and the assumptions built into those models; for example do treated subjects lose all treatment effect immediately or does their disease slowly return? Which assumptions are correct cannot be tested, but sensitivity analyses can be performed to see how model results differ under differing assumptions.

We close this section with some comments about calculations of the standard error of estimates in the PMM framework. If pattern specific estimates for l patterns are obtained, $\{\theta_1, \ldots, \theta_l\}$, they need to be combined to obtain an overall estimate. This is done by weighting the θ_i by the pattern specific proportions $\{p_1, \ldots, p_l\}$ to get $\theta_{PMM} = \sum p_i \theta_i$. This is a simple calculation, however, obtaining the standard error is more difficult because the p_i are estimates themselves and have variability associated with them. We do not illustrate the procedure here but state that the standard error can be calculated by the delta method or by bootstrap (Fairclough 2010). When PMMs are combined with MI, the standard errors are correctly calculated by the MI algorithm without further corrections.

9.14 Discussion

This chapter has been a short synopsis of what we feel is important material on missing data for clinical trials run in the pharmaceutical industry. As a result many topics have been omitted, including a few major topics. One such topic that has been omitted is shared parameter models (Ibrahim and Molenberghs 2009). These models link the longitudinal data model and the missingness model with random effects or latent classes. Although a considerable amount of literature exists, they are rarely used in industry trials. Another topic of interest that has been omitted is double

robust estimation (Kang and Schafer 2007). These methods combine the complete data model and missingness model in such a way that estimates are consistent if either model but non necessarily both are correct. This is an interesting area with a great deal of research activity, but has not been regularly incorporated into pharmaceutical trials as of the writing of this chapter. A related topic that was not covered in this chapter is the use of causal estimands, which focus on the target of the estimation process. That is, what is the θ we are trying to consistently estimate.

We close by encouraging the interested reader to learn more about missing data methods. An area of great importance when analyzing clinical trials. We hope the simple framework of complete data, missingness and analysis models will serve the reader well.

References

Cao, W., Tsiatis, A. A., & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* (Oxford Academic).

Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*. Statistics in Practice. Wiley.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

Fairclough, D. L. (2010). *Design and analysis of quality of life studies in clinical trials*, 2nd ed. Chapman and Hall/CRC.

Ibrahim, J.G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. *Test*, *18*(1), 1–43.

Kang, J. D. Y., & Schafer, L. J. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539.

Mallinckrodt, C. H., Scott Clark, W., & David, S. R. (2001). Type i error rates from mixed effects model repeated measures versus fixed effects ANOVA with missing values imputed via last observation carried forward. *Drug Information Journal, 35*(4), 1215–1225. https://doi.org/10.1177/009286150103500418

Mallinckrodt, C. H., Lin, Q., Lipkovich, I., & Molenberghs, G. (2012). A structured approach to choosing estimands and estimators in longitudinal clinical trials. *Pharmaceutical Statistics*, *11*(6), 456–461.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer.

National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*.

O'Neill, R. T., & Temple, R. (2012). The prevention and treatment of missing data in clinical trials: An FDA perspective on the importance of dealing with it. *Clinical Pharmacology and Therapeutics*, *91*(3), 550–554.

Rubin, D. B. (1996). Multiple imputation after 18 years. *Journal of the American Statistical Association*, *91*(434), 473.

SAS/STAT 14.1 user's guide, Carey, NC (2015).

Chapter 10
Bayesian Subgroup Analysis with Hierarchical Models

Gene Pennello and Mark Rothmann

10.1 Introduction

In studies of investigational treatments, patients are frequently heterogeneous in demographics, disease characteristics, biomarkers, or other variables that are potentially prognostic for a clinical outcome of interest or predictive of the treatment effect on the outcome. Thus, a common practice is to examine if the treatment effect varies in subgroups defined by such variables (Alosh et al. 2015). However, if the treatment is ineffective, and each subgroup is tested for significance at level α ($=0.05$, say), then the probability of observing one or more falsely significant effects within the subgroups (familywise error rate) can be much higher than α. This objection has led to the development of significance testing procedures that control the familywise error rate at α. However, if the treatment effects are heterogeneous among the subgroups, a familywise procedure lowers what is already usually inadequate power to detect significant effects within subgroups due to small sample size.

In addition, due to their uncertainty, the sample estimates of the treatment effects will tend to have too much variation relative to the treatment effects themselves. Thus a sample estimate of the treatment effect within a subgroup may be a random high, that is, large in magnitude when in fact the treatment effect itself is small or null. In summary, separate analyses of treatment effects within subgroups are difficult to interpret for statistical and clinical significance.

G. Pennello (✉)
Division of Biostatistics, Food and Drug Administration, Center for Devices and Radiological Health, 10903 New Hampshire Ave., Silver Spring, MD 20993, USA
e-mail: Gene.Pennello@fda.hhs.gov

M. Rothmann
Division of Biometrics II, Food and Drug Administration, Center for Drug Evaluation and Research, 10903 New Hampshire Ave., Silver Spring, MD 20993, USA
e-mail: Mark.Rothmann@fda.hhs.gov

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_10

Table 10.1 LIFE study summary data

Population	Treatment	N	Person-years[a]	Events	Event percent	Event rate[b]
Overall	Atenolol	4588	21090.4	588	12.82	27.88
	Losartan	4605	21344.5	508	11.03	23.80
Non-Black	Atenolol	4325	19975.0	559	12.92	27.98
	Losartan	4335	20249.3	462	10.66	22.82
Black	Atenolol	263	1115.4	29	11.03	26.00
	Losartan	270	1095.2	46	17.04	42.00

[a]Calculated from label information (US FDA 2003a, b, c), subject to rounding error
[b]per 1000 person-years

To address the difficulty with interpreting separate analyses of subgroups, we consider Bayesian subgroup analysis with hierarchical models. In such models, a prior distribution is placed on the subgroup treatment effects according to an assumed exchangeability structure implemented with random effect distributions. The resulting posterior mean of a subgroup treatment effect borrows strength from all the data, giving it increased precision (as measured by the posterior standard deviation) relative to the sample estimate for that subgroup. In a one-way exchangeability structure, a subgroup treatment effect has a posterior mean that shrinks the sample estimate toward the overall estimate in a weighted averaging of the two quantities. The amount of shrinkage increases with decreased evidence of treatment effect heterogeneity, as measured by the variation between relative to within the subgroups.

Unlike the sample estimates, the shrinkage estimates of the subgroup treatment effects do not (in expectation) have more variation than the treatment effects themselves and are more precise. Therefore, their values may be more clinically interpretable than very high or very low sample estimates that may be due to high imprecision or multiplicity.

Moreover, the amount of shrinkage can vary from subgroup to subgroup. Thus the rank order of the posterior means may be different than those of the sample estimates and may more accurately reflect the true rank order of the treatment effects (Efron and Morris 1975).

10.2 Example

Consider the LIFE study of patients with hypertension, in which losartan was compared with atenolol on time to first major adverse cardiac event (MACE) (US FDA 2003a, b, c). A total of 9193 subjects were randomized to losartan (4605) or atenolol (4588). Most subjects (80%) were European Caucasians between 55 and 80 years old. A total of 1096 MACE events was observed, 508 and 588 in the losartan and atenolol arms, respectively, among 21344.5 and 21090.4 person-years of follow-up (Table 10.1).

Table 10.2 Hazard Ratio (HR) analyses, time to first MACE, Losartan versus Atenolol

Population	Analysis[a]	logHR[b]	SE/SD	HR	95% CI[c]	*p*-value[d]	Pr HR > 1[e]
Overall	F	−0.14	0.06	0.87	0.77, 0.98	0.021	
Non-black	F	−0.19	0.06	0.83	0.73, 0.94	0.003	
	EB	−0.18	0.06	0.83	0.74, 0.93		0.001
	B	−0.18	0.06	0.83	0.74, 0.94		0.002
	FB	−0.20	0.06	0.82	0.72, 0.92		0.001
Black	F	0.51	0.24	1.67	1.04, 2.66	0.033	
	EB	0.43	0.21	1.53	1.01, 2.33		0.978
	B	0.38	0.27	1.46	0.87, 2.46		0.914
	FB	0.39	0.24	1.52	0.93, 1.47		0.948

SE standard error
SD posterior standard deviation
[a]*F* frequentist (unadjusted for multiplicity), *EB* empirical Bayes (Sect. 10.4), *B* more fully Bayes (Sect. 10.5), *FB* fully Bayes for event rate ratio ≈ hazard ratio (Sect. 10.7)
[b]For the frequentist analysis, log HR is the Cox model estimate adjusted for covariates Cornell product, Sokolow-Lyon voltage, and Framingham score. The empirical Bayes and Bayesian analyses were based on these estimates and their standard errors
[c]For frequentist and Bayesian analyses, 95% CI is respectively the 95% confidence interval and the 95% central posterior credible interval
[d]Two-sided *p*-value
[e]Posterior probability that HR > 1

In the overall frequentist analysis, the covariate-adjusted hazard ratio (HR) comparing losartan with atenolol on MACE was 0.87 with 95% confidence interval (CI) (0.77, 0.98) and two-sided *p*-value 0.021, indicating that at level 0.025 losartan was significantly more effective than atenolol at lowering MACE risk (Table 10.2). The same analysis was considered within race subgroups. For non-blacks, HR was 0.83 with 95% CI (0.73, 0.94) and *p*-value 0.003, consistent with the overall results. However, for blacks, HR was 1.67 with 95% CI (1.04, 2.66) and *p*-value 0.033, suggesting that losartan is worse than atenolol at lowering MACE risk in this subpopulation. Is this finding real?

10.3 Bayesian Hierarchical Modeling for Subgroup Analysis

We now describe general assumptions for a Bayesian hierarchical model using normal data. We provide some analytical results on Bayesian posterior distributions, interval estimates and hypothesis tests for subgroup analysis given that the parameters of the model are known.

For simplicity, we assume that for subgroups $j = 1, 2, \ldots, J$ we can by sufficiency reduce the data to sample estimates $\underline{y} = (y_1, \ldots, y_J)$ of the treatment effects $\underline{\mu} = (\mu_1, \ldots, \mu_J)$ given that the variances $\underline{\sigma}^2 = \left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_J^2\right)$ of $\underline{y}$ are known. The sample estimates are assumed independent with

$$y_j|\mu_j \sim N\left(\mu_j, \sigma_j^2\right), \tag{10.1}$$

$j = 1, 2, \ldots, J$. In a Bayesian hierarchical model, a prior distribution is placed on the treatment effects $\underline{\mu}$. Specifically, the treatment effects are assumed independent with

$$\mu_j \sim N\left(\mu_0, \sigma_\mu^2\right) \tag{10.2}$$

where μ_0 is the mean and σ_μ^2 is the between-subgroup variance of the treatment effects. Because the treatment effects are independent and identically distributed in their prior distribution, they are exchangeable, that is, any ordering of their values is equally plausible a priori.

Posterior Distribution. Assume the values of μ_0, σ_μ^2, and $\underline{\sigma}^2 = \left(\sigma_j^2, j = 1, 2, \ldots, J\right)$ are known. Let $\underline{\theta}_j = \left(\mu_0, \sigma_\mu^2, \sigma_j^2\right)$. Then for subgroup j the posterior distribution of the treatment effect μ_j is given by

$$\begin{aligned} \mu_j|\underline{y}, \underline{\theta} &\sim N\left(\left(1 - S_j\right)\mu_0 + S_j y_j, S_j\sigma_j^2\right) \\ &\equiv N\left(E_j, V_j\right) \end{aligned} \tag{10.3}$$

where $S_j = 1 - \Phi_j^{-1}$ is a shrinkage factor and $\Phi_j = (\sigma_j^2 + \sigma_\mu^2)/\sigma_j^2$ the "true" F ratio for subgroup j. The posterior mean $E_j = \left(1 - S_j\right)\mu_0 + S_j y_j$ is a weighted average of the sample estimate y_j and the prior mean μ_0. As the variation σ_μ^2 between the subgroups decreases, the weight S_j on y_j decreases in favor of more weight on mean μ_0.

Hypothesis Testing. In Bayesian hypothesis testing, we conclude $\mu_j > 0$ if posterior probability

$$Pr(\mu_j > 0|\underline{y}, \underline{\theta}_j) > 1 - \alpha$$

for some $\alpha \in (0, 0.5)$. For example, if $\alpha = 0.025$. we conclude $\mu_j > 0$ if its posterior probability is greater than 0.975. From the posterior distribution (10.3),

$$\begin{aligned} &Pr\left(\mu_j > 0|\underline{y}, \underline{\theta}_j\right) \\ &= Pr\left(Z > -E_j/\sqrt{V_j}|\underline{y}, \underline{\theta}_j\right) \\ &= \varphi\left(E_j/\sqrt{V_j}\right) \end{aligned} \tag{10.4}$$

where $Z \sim N(0, 1)$ is a standard normal variable with cumulative distribution function $\varphi(\bullet)$.

Interval Estimation. From 10.3, a $1 - \alpha$ Bayesian credible interval on μ_j is

$$E_j \pm z_{1-\alpha/2}\sqrt{V_j} \tag{10.5}$$

where $\varphi(z_\gamma) = \gamma, 0 < \gamma < 1$. That is, the posterior probability that μ_j is in the interval (10.5) is $1 - \alpha$. In symbols,

$$Pr(\mu_j \in \left[E_j \pm z_{1-\alpha/2}\sqrt{V_j}\right] | \underline{y}, \underline{\theta}_j) = 1 - \alpha$$

10.4 Empirical Bayes Subgroup Analysis

In an empirical Bayes analysis, the parameters of the model—mean treatment effect μ_0, between subgroup variance σ_μ^2, and within-subgroup variances $\underline{\sigma}^2$—are replaced by empirical estimates. For example, in the LIFE study, σ_j^2 can be set to the square of the standard error on the log hazard ratio y_j. for subgroup j, $j = 1, 2, \ldots, J$. Given these estimates are substituted for $\underline{\sigma}^2$, we now describe weighted least squares (WLS) estimators of μ_0 and σ_μ^2. From 10.1 and 10.2, the marginal distribution of the sample estimate y_j is given by

$$y_j \sim N\left(\mu_0, w_j^{-1}\right),$$

where $w_j = 1/\left(\sigma_j^2 + \sigma_\mu^2\right)$. For this marginal model, the mean square error is

$$MSE\left(\sigma_\mu^2\right) = (J-1)^{-1} \sum_{j=1}^{J} w_j \left(y_j - \hat{\mu}_0\left(\sigma_\mu^2\right)\right)^2$$

where

$$\hat{\mu}_0\left(\sigma_\mu^2\right) = \left(\sum_{j=1}^{J} w_j\right)^{-1} \sum_{j=1}^{J} w_j y_j.$$

For all values of $\underline{\sigma}^2$ and σ_μ^2, $MSE\left(\sigma_\mu^2\right)$ has expectation 1. Thus, given $\underline{\sigma}^2$, the WLS estimates of σ_μ^2 and μ_0 are $\hat{\sigma}_\mu^2$ and $\hat{\mu}_0\left(\hat{\sigma}_\mu^2\right)$ where $MSE\left(\hat{\sigma}_\mu^2\right) = 1$.

Life Study Results. To illustrate, for the LIFE study, the log hazard ratios were $y_1 = -0.19$ and $y_2 = 0.51$ for the non-black and black subgroups (Table 10.2). Asymptotically, they follow a normal distribution, permitting analysis under the model defined by 10.1 and 10.2. The sample variances for the log hazard ratios

were $\sigma_1^2 = 0.0036$ and $\sigma_2^2 = 0.0576$, the squares of the standard errors. Given these values, the weighted least squares estimates are $\hat{\mu}_0 = 0.1214$ for the mean and $\hat{\sigma}_\mu^2 = 0.2144$ for the between subgroup variance. (see Appendix, Matlab code, MathWorks® 2014). Thus, the estimates of the "true" F ratios are $\hat{\Phi}_1 = (\sigma_1^2 + \hat{\sigma}_\mu^2)/\sigma_1^2 = (0.0036 + 0.2144)/0.0036 = 60.6$ and $\hat{\Phi}_2 = (\sigma_2^2 + \hat{\sigma}_\mu^2)/\sigma_2^2 = (0.0576 + 0.2144)/0.0576 = 4.72$. Accordingly, the estimates of the shrinkage factors are $\hat{S}_1 = 1 - \hat{\Phi}_1^{-1} = 0.9835$ and $\hat{S}_2 = 1 - \hat{\Phi}_2^{-1} = 0.7882$.

Using the weighted average formula for posterior mean E_j given in (10.3), the posterior mean of the log hazard ratio is $E_2 = 0.43$ for blacks, considerably smaller than the sample estimate $y_2 = 0.51$. Using the formula for V_j, the corresponding posterior standard deviation (SD) is 0.2131. The hazard ratio estimate is $exp(E_2) = 1.53$. The 95% Bayesian credible interval is $exp(0.43 \pm 1.96(0.2131)) = exp(0.0101, 0.8453) = (1.01, 2.33)$. The posterior probability that the hazard ratio is greater than 1 is $\varphi(0.43/0.2131) = 0.978$. In this analysis, MACE risk is concluded to be larger for black patients on losartan than those on atenolol. However, the uncertainty of the estimates for the mean $\hat{\mu}_0$ and the between subgroup variance $\hat{\sigma}_\mu^2$ is not being considered. These estimates are highly uncertain, especially when considering that they estimate the parameters of the prior distribution (10.2) based on the data for just two subgroups.

10.5 A More Fully Bayes Subgroup Analysis

In a more fully Bayesian analysis of the LIFE study, μ_0 and σ_μ^2 are given diffuse prior distributions to estimate them essentially from the data with a posterior distribution that accounts for estimation uncertainty. To this end, we assume the prior distributions for μ_0 and σ_μ^2 are independent with

$$\mu_0 \sim N(0, 16), \text{ and}$$
$$\sigma_\mu^{-2} \sim \Gamma(0.001, 0.001).$$

Under these priors, obtaining the posterior distribution of $\mu_j|\underline{y}, \underline{\sigma}^2$ is difficult analytically, involving integration of the posterior distribution (10.3) over the appropriate conditional distribution of μ_0 and σ_μ^2. Instead, we used Gibbs sampling to sample parameter values of μ_j, μ_0, and σ_μ^2 (Tanner 1996). The sampled values of μ_j converge to its posterior distribution. We implemented the Gibbs sampler with code (Appendix) written for the software package OpenBUGS (Lunn et al. 2000).

As before, we have assumed that the within subgroup variances $\underline{\sigma}^2$ are known and equal to the sample variances. If uncertainty in estimating $\underline{\sigma}^2$ is a concern, then the likelihood for the sample variances could be used for analysis in combination with a diffuse prior on $\underline{\sigma}^2$ to obtain the posterior distribution of $\mu_j|\underline{y}$ that accounts for estimation uncertainty of all the parameters in a fully Bayes analysis.

Life Study Results. For blacks, the posterior mean of the log hazard ratio is 0.38, considerably smaller than the sample estimate $y_2 = 0.51$ (Table 10.2). The cor-

responding posterior SD of 0.27 reflects more uncertainty than its empirical Bayes counterpart (0.21). The hazard ratio estimate is $exp(0.38) = 1.46$. The 95% Bayesian credible interval, given by the 2.5 and 97.5 percentiles of the posterior distribution, is $(0.87,\ 2.46)$. (In general, this central posterior interval is different from the highest posterior density interval, which can be shorter.)

The posterior probability that the hazard ratio is greater than 1, obtained by monitoring the Gibbs samples for $I\left(\mu_j > 0\right)$, is 0.914 which is, smaller than empirical Bayes estimate 0.978. Thus, the possibility is left open that the observation in blacks that MACE risk was higher for losartan than atenolol could have been due to chance. Still, among blacks the probability still comfortably favors a larger treatment effect for atenolol than for losartan.

10.6 Difference in Treatment Effect Between Subgroups

Consider the posterior distribution for the difference in treatment effects $\delta_{12} = \mu_1 - \mu_2$ between subgroups 1 and 2 given the parameters $\underline{\theta}_{12} = \left(\mu_0,\ \sigma_\mu^2,\ \sigma_1^2,\ \sigma_2^2\right)$. From Eq. 10.3,

$$\delta_{12}|\underline{y},\ \underline{\theta}_{12} \ \sim\ N\left((S_2 - S_1)\mu_0 + S_1 y_1 - S_2 y_2,\ S_1\sigma_1^2 + S_2\sigma_2^2\right) \tag{10.6}$$

In the balanced case when $\sigma_j^2 \equiv \sigma_y^2$ for all $j = 1,\ 2,\ \ldots,\ J$, $S_j \equiv S$, and the posterior distribution reduces to

$$\begin{aligned} \delta_{12}|\underline{y},\ \underline{\theta}_{12} \ &\sim\ N\left(Sd_{12},\ S\sigma_d^2\right) \\ &\equiv N(E_{12},\ V_{12}), \end{aligned} \tag{10.7}$$

where $d_{12} = y_1 - y_2$ is the difference between sample treatment effects with variance $\sigma_d^2 = 2\sigma_y^2$. Note the posterior mean of δ_{12} shrinks the sample difference d_{12} toward 0 by the shrinkage factor S.

Hypothesis Testing. In Bayesian hypothesis testing, $\delta_{12} > 0$ is concluded when

$$\Pr(\delta_{12} > 0|\underline{y},\ \underline{\theta}_{12}) > 1 - \alpha$$

for some $\alpha \in (0,\ 0.5)$. From 10.7,

$$\begin{aligned} &\Pr\left(\delta_{12} > 0|\underline{y}, \underline{\theta}_{12}\right) \\ &= \Pr\left(Z > -E_{12}/\sqrt{V_{12}}|\underline{y}, \underline{\theta}_{12}\right) \\ &= \varphi\left(E_{12}/\sqrt{V_{12}}\right) \\ &= \varphi\left(z_{12}\sqrt{S}\right) \end{aligned}$$

where $z_{12} = d_{12}/\sigma_d$ is the standardized difference. Thus $\delta_{12} > 0$ is concluded when

$$z_{12} > \frac{z_{1-\alpha}}{\sqrt{S}}.$$

where $z_{1-\alpha}$ is the nominal α level critical value ($1-\alpha$ th quantile) of the standard normal distribution. Notice, the Bayesian critical value $z_{1-\alpha}/\sqrt{S}$ for declaring that standardized difference z_{12} is "significant" is always larger than the nominal value $z_{1-\alpha}$ unless $\sigma_\mu^2 = \infty$ ($S = 1$) and increases as ratio σ_μ^2/σ_y^2 of the variation between to the variation within subgroups decreases. In a more fully Bayesian analysis in which μ_0 and σ_μ^2 are unknown and given diffuse prior distributions, the critical value can be less than the nominal critical value when the variation between relative to within the subgroups is large enough (Waller and Duncan 1969, 1972).

Interval Estimation. From 10.7, a $1-\alpha$ Bayesian credible interval on δ_{12} is

$$Sd_{12} \pm z_{1-\alpha/2}\sigma_d\sqrt{S} \tag{10.8}$$

Because $S \leq 1$, the half-width of this interval, $z_{1-\alpha/2}\sigma_d\sqrt{S}$, is no wider and perhaps considerably narrower than the half-width $z_{1-\alpha/2}\sigma_d$ of the nominal $1-\alpha$ frequentist confidence interval. The narrower width is due to assuming in the prior that treatment effects within subgroups are exchangeable, which enables information borrowing, which in turn increases precision of estimation relative to the sample estimates within subgroups.

Life Study Results. For the LIFE study, the ratio of the black hazard ratio to the non-black hazard ratio comparing losartan with atenolol has from 10.6 empirical Bayes posterior mean $exp(E_{21}) = \exp(0.6126) = 1.8452$, which is shrunk relative to the sample estimate $exp(0.51 + 0.19) = 2.01$. From the delta method the corresponding posterior SD is $exp(E_{21})\sqrt{V_{21}} = 1.8452(0.2212) = 0.4082$, approximately. The 95% credible interval is $\exp(0.6126 \pm 1.96(0.2212)) = (1.196, 2.847)$. The posterior probability that ratio of the hazard ratios is greater than 1 is $\varphi(0.6126/0.2212) = \varphi(2.769) = 0.9972$, indicating that the hazard ratios are quantitatively different, according to this analysis.

10.7 Effect Modifiers in Subgroup Analysis

A criticism of univariate subgroup analyses is that each factor that defines a set of subgroups is evaluated in isolation, with potential effect modification of the other factors ignored (Varadhan and Wang 2014). Generally, a treatment by covariate interaction may be due to the covariate being correlated with one or more treatment effect modifiers (variables causally related with treatment effect size) even while the covariate itself is not a treatment effect modifier. Without adjustment for the covariate (or the effect modifiers), resulting inference on the treatment effect may be underpowered, biased, or both (Senn 1989). Generalizing, if several factors contribute small

or moderate effect modification, a marginal evaluation of heterogeneous treatment effects across the levels of one of the factors in isolation of the others could be underpowered or biased. This point is important as an individual patient may have a collection of characteristics that are individually correlated with treatment effect. However, evaluating the treatment effect within many potential treatment effect modifiers simultaneously would require including them all in a model along with corresponding interactions with treatment. When evaluating a treatment effect within levels of covariates suspected or known to be treatment effect modifiers, a reasonable assumption is that the treatment by covariate interactions are exchangeable (Dixon and Simon 1991, 1992; Simon 2002).

To illustrate modeling of random interactions with treatment, we implement a fully Bayesian analysis of the LIFE study incident rate $\lambda_{aj} = \exp(\mu_{aj})$ of MACE per person-year for treatment arms atenolol and losartan $a = 1, 2$ within non-black or black race subgroups $j = 1, 2$. For small event rates, the rate ratio $\rho_j = \lambda_{2j}/\lambda_{1j}$ is approximately equal to the hazard ratio when hazards are proportional (Holford 1980; Jewell 2004, p. 36, Eq. 4.6).

For atenolol and losartan, the MACE event counts $\{x_{aj}\}$ were $x_{11} = 559$ and $x_{21} = 462$ in non-blacks and $x_{12} = 29$ and $x_{22} = 46$ in blacks. Corresponding person-years were $y_{aj} = 19975.0, 20249.3, 1115.4, 1095.2$, accumulated for patient samples sizes $n_{aj} = 4325, 4335, 263, 270$ (Table 10.1). Estimates are $\hat{\lambda}_{aj} = x_{aj}/y_{aj}$ for the MACE rate per person-year and $\hat{\mu}_{aj} = \log\left(\hat{\lambda}_{aj}\right)$ for the log rate.

In a quasi-likelihood analysis based on the Poisson distribution, an event count x has mean and variance

$$Ex = y\lambda, \quad Vx = \phi y\lambda,$$

where y is the number of person-years and ϕ is potential over- or under-dispersion relative to Poisson variation. Under this model,

$$\hat{\mu}_{aj} \sim N(\mu_{aj}, \phi_{aj}/x_{aj})$$

approximately. If person-years y_{ajk} and events x_{ajk} were available for each individual patient $k = 1, 2, \ldots, n_{aj}$, then ϕ_{aj} may be estimated as $\hat{\phi}_{aj} = S^2_{aj}/f_{aj}$, where $S^2_{aj} = f^{-1}_{aj} \sum_{k=1}^{n_{aj}} y_{ajk}\left(\hat{\lambda}_{ajk} - \hat{\lambda}_{aj}\right)^2$ is a weighted sum of squares over the replications with $f_{aj} = n_{aj} - 1$ degrees of freedom and $\hat{\lambda}_{ajk} = x_{ajk}/y_{ajk}$ (McCullagh and Nelder 1989, Sect. 4.5.2). The statistic S^2_{aj} is independent of $\hat{\mu}_{aj}$ with data distribution given by

$$S^2_{aj}/\phi_{aj} \sim \chi^2(f_{aj})$$

approximately. Because we did not have individual patient data, for the sake of analysis we presumed $\hat{\phi}_{aj} = 1$ for all $a, j = 1, 2$.

We assume that μ_{aj} is linear in the predictors. Specifically,

$$\mu_{aj} = \tau_a + \beta_j + \gamma_{aj}$$

where $\{\tau_a\}$ are fixed treatment effects, $\{\beta_j\}$ are random race effects, and $\{\gamma_{aj}\}$ are random treatment by race interaction effects. We assume the random effect distributions are given by

$$\beta_j \sim N\left(0, \sigma_\beta^2\right) \gamma_{aj} \sim N\left(0, \sigma_\gamma^2\right)$$

To implement a fully Bayesian analysis, we place independent, diffuse prior distributions on the unknown parameters:

$$\begin{aligned} \tau_a &\sim N(0, 1000), \\ 1/\sigma_\beta^2,\ 1/\sigma_\gamma^2,\ 1/\phi_{aj} &\sim \Gamma(0.01, 0.01). \end{aligned}$$

The code for implementation in OPENBugs is given in the Appendix.

Life Study Results. For blacks, the Bayes posterior mean of log rate ratio is 0.39 (Table 10.2). The 95% Bayesian central posterior credible interval is (0.93, 1.47). The posterior probability that the rate ratio is greater than 1 is 0.948, leaving open the possibility that in black patients MACE rate may in fact not be higher for losartan than atenolol, although the rate ratio may be no more smaller than 0.93 in favor of losartan.

10.8 Multi-way Bayesian Subgroup Analysis

For two or more factors that define sets of subgroups (e.g., age group, sex, race/ethnicity, region), hierarchical modeling can be extended to accommodate an exchangeability structure for the factors. For example, in a one-way analysis, treatment effects within combinations of race and region are considered completely exchangeable, with the structure of the factors ignored. Alternatively, the factors may be additive or may interact in their modification of treatment effect. For race and region factors A and B, a two-way model can be used with main effects of race (A), main effects of region (B) and interaction effects of race by region (C) each separately considered exchangeable. As shown below, the Bayesian posterior mean of the difference in treatment effect between subgroups defined by factor A is an intuitive linear combination of marginal and interaction contrasts that are shrunk according to evidence for main A effects and interaction C effects. As an individual patient belongs to many subgroups, these more complex models can be used to estimate the treatment effect an individual patient can expect.

To illustrate the behavior of multi-way hierarchical models, we briefly describe some analytical results for the two-way hierarchical model of normal data assuming the parameters of the model are known.

Consider two separate factors A and B having levels $i = 1, \ldots, I$ and $j = 1, \ldots, J$ in a balanced design of normally distributed sample treatment effects $\underline{y} = \{y_{ij},\ i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, J\}$ with homogeneous error variance σ_y^2. A two-way Bayesian hierarchical model is

$$\begin{aligned} y_{ij} &\sim N(\mu_{ij}, \sigma_y^2) \\ \mu_{ij} &= \mu_0 + \alpha_i + \beta_j + \gamma_{ij} \end{aligned}$$

with independent priors

$$\alpha_i \sim N(0, \sigma_\alpha^2), \quad \beta_j \sim N(0, \sigma_\beta^2), \quad \gamma_{ij} \sim N(0, \sigma_\delta^2), \quad \mu_0 \sim N(\theta, \sigma_\theta^2)$$

and $\left(\sigma_y^2, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\delta^2\right)$ assumed known. According to Pennello (1997), under this model the posterior distribution of the difference in treatment effect

$$\delta_{12,j} = \mu_{1j} - \mu_{2j}$$

between levels 1 and 2 within subgroup j is given by

$$\delta_{12,j} | \underline{y}, \underline{\boldsymbol{\sigma}}^2 \sim N\left(S_A \bar{d}_{12,\bullet} + S_C d_C, (S_A + (J-1)S_C)\sigma_d^2/J\right)$$

where

$$\begin{aligned} \bar{d}_{12,\bullet} &= \bar{y}_{1\bullet} - \bar{y}_{2\bullet}, \quad d_C = d_{12,j} - \bar{d}_{12,\bullet}, \quad d_{12,j} = y_{1j} - y_{2j}, \sigma_d^2 = 2\sigma_y^2, \\ S_A &= 1 - 1/\Phi_A, \quad \Phi_A = \sigma_A^2/\sigma_y^2, \quad \sigma_A^2 = \sigma_C^2 + J\sigma_\alpha^2 \\ S_C &= 1 - 1/\Phi_C, \quad \Phi_C = \sigma_C^2/\sigma_y^2, \quad \sigma_C^2 = \sigma_y^2 + \sigma_\delta^2 \end{aligned}$$

In Bayesian hypothesis testing, $\delta_{12,j} > 0$ is concluded if

$$\Pr(\delta_{12,j} > 0 | \underline{y}, \underline{\boldsymbol{\sigma}}^2) > 1 - \alpha$$

for some $\alpha \in (0,\ 0.5)$. Equivalently, $\delta_{12,j} > 0$ is concluded if

$$z_{12,j} > \frac{z_{1-\alpha}}{\sqrt{S_C}} \left\{ \frac{S_A}{J S_C} + \frac{J-1}{J} \right\} - \frac{\bar{z}_{12,\bullet}}{\sqrt{J}} \left\{ \frac{S_A}{S_C} - 1 \right\} \tag{10.9}$$

where

$$z_{12,j} = d_{12,j}/\sigma_d,$$

is the standardized difference between levels 1 and 2 within subgroup j and

$$\bar{z}_{12,\bullet} = \bar{d}_{12,\bullet} / \left(\sigma_d/\sqrt{J}\right)$$

is the marginal standardized difference.

The Bayesian critical value for $z_{12,j}$ (right-hand term in Eq. 10.9) decreases linearly as marginal standardized difference $\bar{z}_{12,\bullet}$ increases. The coefficient on $\bar{z}_{12,\bullet}$ increases as the ratio S_A/S_C of the shrinkage factors for the marginal A and interaction C effects increase, indicating less evidence for interaction relative to main effects, and decreases with the number of subgroups J.

In the limit as the marginal main A effect variance σ_α^2 tends toward zero, the posterior distribution reduces to

$$\delta_{12,j}|\underline{\boldsymbol{y}},\ \underline{\boldsymbol{\sigma}}^2 \sim N\left(S_C d_{12,j},\ S_C \sigma_d^2\right)$$

which corresponds to the posterior distribution for the difference in a one-way model of sample effects $\left\{y_{ij},\ i = 1,\ 2,\ \ldots,\ I\right\}$. Thus, the one-way model in Sect. 10.3 can be seen to be an approximation to the two-way model if evidence for interaction effects dominates evidence for main effects.

In a fully Bayesian analysis, the likelihood for σ^2 may be utilized in combination with Jeffreys prior on $\left(\mu_0, \sigma^2,\ \sigma_\alpha^2,\ \sigma_\beta^2,\ \sigma_\delta^2\right)$ (Box and Tiao 1973) to obtain the posterior distribution of $\delta_{12,j}|\underline{\boldsymbol{y}}$ that fully accounts for estimation uncertainty for these variance components (Pennello 1997).

10.9 Discussion

Bayesian subgroup analysis offers efficiency in estimation and generally provides more precise point estimates and narrower interval estimates than standard frequentist analyses unadjusted for multiplicity. Increased precision is possible because subgroups are analyzed jointly rather than separately, invoking information borrowing. Treatment effects within subgroups are assumed exchangeable (random with a common distribution), allowing the treatment effects to be different, but related. When estimating the effect in a given subgroup, the outcomes of all patients are considered relevant, more so for those subjects within the subgroup of interest.

Using a Bayesian hierarchical model, a sample estimate of a subgroup treatment effect that is large and clinically impressive in magnitude could be shrunk to a much smaller, less compelling value. Such dramatic shrinkage could suggest that the large point estimate may have been a random high due to multiple testing and large standard errors within subgroups. Additionally, we can account for the correlation structure of covariates that are effect modifiers by using a multi-way hierarchical model.

Bayesian hierarchical models enjoy good frequentist properties. In particular, under a loss function in which the losses for making directional decisions on the treatment effects within subgroups are additive and the loss is 0, 1, and $A (0 < A < 1)$ for correct, incorrect, and non-committal decisions on treatment effect direction, the Bayes rule for the one-way hierarchical model controls the directional false discovery rate at A (Lewis and Thayer 2004).

In one-way or multi-way random effects models with unknown variance components, Gibbs samples of the parameters may mix poorly (be highly autocorrelated), resulting in slow convergence of the samples to the posterior distribution. Strategies for improving mixing include reparameterizations that are hierarchically centered (Gelfand et al. 1996) or that sweep random effects means to lower order terms (Gilks and Roberts 1996). We employed hierarchical centering in our Gibbs sampling code of the one-way model (Appendix), which helped to accelerate convergence. As explained by the references just cited, without such reparameterizations of random effect models, Gibbs samples of some parameters will be highly auto-correlated if the sample error variances are small relative to the between effect variances, which was the case for the LIFE study (for log hazard ratio, WLS estimate of between subgroup variance 0.2144 was large compared with error variances 0.0036 and 0.0576 for non-blacks and blacks).

Gilks et al. (1996) provide an excellent introduction to Markov chain Monte Carlo methods, including Gibbs sampling. Tanner (1996) provides many worked examples of how to construct Gibbs sampling algorithms for frequently encountered data models and prior distributions. For survival models, Kuo and Smith (1992) provide useful Gibbs sampling algorithms. For implementation of the counting process approach for modeling baseline hazard and regression parameters, see the example Leuk in the OPENBugs package.

To promote the use of Bayesian methods for subgroup analysis in patient-centered outcomes research, web-based software tools have been developed (Henderson et al. 2016; Wang et al. 2018). Such software is designed to lower barriers in implementing Bayesian subgroup analysis.

For a small number of subgroups, inferences may be sensitive to the prior placed on the variance between subgroups in the treatment effect. For a one-way model of subgroup-specific treatment effects in a cross-over design structure, Hsu et al. (2017) compare the posterior distributions of the treatment effects under several priors for the between subgroup variance. They also obtain the posterior distribution for the subgroup with the smallest treatment effect in an assessment of treatment effect consistency across subgroups. Similarly, the largest mean problem has been considered in a Bayesian decision theoretic framework of a one-way hierarchical model (Bland and Duncan 1964).

In the LIFE study, the validity of the shrinkage estimates for the hazard ratios among blacks and non-blacks is predicated on exchangeability of the subgroups, that is, not expecting a worse treatment effect in blacks than in non-blacks a priori. A more flexible model than complete exchangeability of the treatment effects considered here is to place a Dirichlet prior on the distribution of the treatment effects. Such "non-parametric" Bayesian analyses do not force sample estimates to shrink excessively when similarity of treatment effects is not supported but can still result in greater precision (narrower credible intervals) than separate analyses of the subgroups (Gamalo-Siebers et al. 2016).

Exchangeability of treatment effects within subgroups may not be reasonable a priori. For example, for a binary biomarker that is the target of a treatment, the treatment effect may a priori be expected to be greater in subjects who are biomarker

positive than those who are biomarker negative. Prior distributions have been developed that are tailored to treatment effects within subgroups defined by predictive biomarkers (Karuri and Simon 2012).

Appendix

Section 10.4 Analysis. Matlab code (MathWorks® 2014) for empirical Bayes analysis for LIFE Study log hazard ratios.

```
lhr=[-0.19; 0.51];          % log hazard ratios
se2=[0.0036; 0.0576];       % squared standard errors
z=-abs(lhr)./sqrt(se2)      % standardized z value
pr=normcdf(z)               % 1-sided p value for null HR = 1

the=[lhr' se2'];            % parameter
mseminus1 = @(s2mu) msewls(the(1:2), the(3:4), s2mu) - 1; % parameterized function
s2mu=fzero(mseminus1, 0.1) % WLS estimate of between subgroup variance s2mu=0.2144
mu0=mu0wls(lhr, se2, s2mu) % WLS estimate of mu0

if 0                        % code to check fzero result
mse=msewls(lhr, se2, s2mu) % =1, if fzero result is correct
v=se2+s2mu; w=1./v
lhr0hat=sum(w.*lhr)/sum(w) % should = mu0
end;

Phi=1+s2mu./se2             % F ratios
S=1-1./Phi                  % shrinkage factors
E=(1-S).*mu0 + S.*lhr       % posterior mean
V=S.*se2                    % posterior variances
zB=E./sqrt(V)               % posterior z values
prB=normcdf(zB)             % posterior probability HR > 1

smypr=[pr prB]              % summary of p values, posterior probabilities
prhyp=1-smypr

alp=0.025;  zalp=norminv(1-alp);
cilhr=E*ones(1,2) + zalp*sqrt(V)*[-1 1]
cihr=exp(cilhr)             % credible interval on HR
smy=[exp(E) exp(E).*sqrt(V) cihr]

lam=[-1; 1];                % contrast for difference in log HR between 2 subgroups
lhrdif=lam'*E               % difference in posterior mean log HR between subgroups
lhrdifsd=sqrt(sum(V))       % posterior SD of difference
zlhrdif=lhrdif/lhrdifsd     % posterior z value
cilhrdif=lhrdif*ones(1,2) + zalp*lhrdifsd*[-1 1]
hrdif=exp(lhrdif)           % posterior estimate of ratio of HRs
cihrdif=exp(cilhrdif)       % credible interval for ratio of HRs
smylhrdif=[lhrdif lhrdifsd cilhrdif normcdf(-zlhrdif)]
smyhrdif=[hrdif hrdif*lhrdifsd cihrdif]
```

```
function mse=msewls(y, se2, s2mu);
% function msewls(y, se2, s2mu)
% gets weighted mean square error
% under 1-way random effects model
% of vec y with sample variances se2 and between variance s2mu.
a=length(se2); f=a-1; %o=ones(a,1);
v=se2+s2mu; w=1./v;
mu0=sum(w.*y)/sum(w); %mu0=mu0wls(y, se2, s2mu)
mse=sum(w.*(y-mu0).^2)/f;

function mu0=mu0wls(y, se2, s2mu);
% function mu0wls(y, se2, s2mu)
% gets weighted least squares estimate of
% overall mean under 1-way random effects model
% of vec y with sample variances se2 and between variance s2mu.
a=length(se2); f=a-1; o=ones(a,1);
v=se2+s2mu; w=1./v;
mu0=sum(w.*y)/sum(w);
```

Output of Matlab Analysis

```
smypr =
    0.0008    0.9991
    0.0168    0.0224
prhyp =
    0.9992    0.0009
    0.9832    0.9776
cihr =
    0.7397    0.9340    -0.1849
    1.0101    2.3288     0.4277
smyhr -
    0.8312    0.0495    0.7397    0.9340
    1.5337    0.3268    1.0101    2.3288
hrdif =
    1.8452
cihrdif =
    1.1960    2.8467
smylhrdif =
    0.6126    0.2212    0.1790    1.0462    0.0028
smyhrdif =
    1.8452    0.4082    1.1960    2.8467
```

Section 10.5 Analysis. OpenBUGS code (Lunn et al. 2000) for analysis of LIFE study log hazard ratio mu[s] in one-way normal-normal hierarchical model of sample log hazard ratios sest[s] with variances s2.sest[s] assumed known and equal to square of standard errors for subgroups s = 1, 2.

Model

```
model
{ for(s in 1:S) {                          # S= race subgroup 1, 2
    prec.sest[s] <- 1/s2.sest[s]           # prec[s] = sample precision of sest [s]
    sest[s] ~dnorm(mu[s], prec.sest[s])    # sest[s] = sample estimate of log HR [s]
    mu[s] ~dnorm(mu0, prec.mu)             # mu[s] = log HR are random normal (exchangeable)
    prob[s] <- step(opc - mu[s]);          # prob[s]=probability mu[s] > opc
    rr[s] <- exp(mu[s])                    # rr[s]= HR[s]
}
tau2.mu0 <- 1/var.mu0                      # common variance of log HR[s]
mu0 ~ dnorm(0, tau2.mu0)                   # prior on common mean for log HR[s]
prec.mu ~ dgamma(.001,.001)                # prior on prec.mu= log HR precision subgroups
tau2.mu <- 1/prec.mu                       # variance between subgroups in log HR
 }
```

Data, log hazard ratios

```
list(S=2, sest=c(-0.19, 0.51), s2.sest=c(0.0036, 0.0576 ), var.mu0=16, opc=0)
```

Data, log rate ratios

```
list(S=2, sest=c(-0.2042, 0.4796), s2.sest=c(0.00395, 0.05622 ), var.mu0=16, opc=0)
```

Inits

```
list(mu0=0, prec.mu=1)
```

Section 10.7 Analysis. OpenBUGS code (Lunn et al. 2000) for analysis of LIFE study MACE rate based on quasi-likelihood for canonical Poisson generalized linear model with fixed treatment effects, random race effects, and random treatment by race interaction effects.

Model

```
model
{
        for(r in 1 : R ) {rac[r] ~ dnorm (0, tau[1])   }# rac = race effects, random
        for(t in 1 : T ) { trt[t] ~ dnorm (0, ooo1)   } # trt = treatment effects, fixed
                for( r in 1 : R ) { for(t in 1 : T ) {      # x[r, t]= events for rac r, trt t
                    rate[r, t] <- x[r, t] / py[r, t]        # py[r, t]= person-years for rac r, trt t
                  lograte[r, t] <- log(rate[r, t])
    lograte[r, t] ~ dnorm(mu[r, t], prec.dat[r, t])
   mu.gam[r, t] <- trt[t]+rac[r]
                     mu[r, t] ~ dnorm(mu.gam[r, t], tau[2]) # hierarchical centering of gam
                  gam[r, t] <- mu[r, t]-mu.gam[r, t]  # gam=rac by trt interactions, random
                    lam[r, t] <- exp(mu[r, t])           # lam= rate per py by rac r, trt t
```

```
      prec.dat[r, t] <- x[r, t]/phi[r, t]          # prec.dat[r,t]= precision of lograte[r,t]
           f[r, t] <- n[r, t] - 1                  # f = dof on phihat = n-p (p=1 here)
#phi[r,t]<- 1/prec[r,t]                           # phi = overdispersion
#phisig[r,t]<- sqrt(phi[r,t])                     # phisig= overdispersion sd
#aa[r,t]<- f[r,t]/2                               # phihat[r, t]= X2 or deviance / n - p
#bb[r,t]<- aa[r,t]*prec[r,t]
#phihat[r,t]~dgamma(aa[r,t],bb[r,t])              # phihat[r, t] likelihood
#prec[r,t]~dgamma(o1,o1)                          # prior on 1/phi
          }}

            for(r in 1:2) { rr[r] <- lam[r,2] / lam[r,1]  # rr[r] = rate ratio for rac r
        logrr[r] <- log(rr[r])
                  prb[r] <- step(opc - rr[r]); }  # probability rr[r] > opc
            for(k in 1:2) {
              tau[k] ~ dgamma(o1, o1)      # tau[k] = prec on random effect k
              sig[k] <- 1 / sqrt(tau[k])  }      # sig[k] =   sd on random effect k
}
Data, MACE Event Rate
list(R=2, T=2, o1=0.01, ooo1=0.001, a=0.01, b=0.01, opc=1,
phihat = structure(.Data = c(1,1,1,1), .Dim = c(2,2)),
  n = structure(.Data = c(4325, 4335,  263,  270), .Dim = c(2,2)),
 py = structure(.Data = c(19975,20249.3,1115.4,1095.2), .Dim = c(2,2)),
  x = structure(.Data = c(559,462,29,46), .Dim = c(2,2)) )

Inits
list(tau = c(1,1), prec = structure(.Data = c(1,1,1,1), .Dim = c(2,2)) )
```

References

Alosh, M., Fritsch, K., Huque, M., Mahjoob, K., Pennello, G., Rothmann, M., et al. (2015). Statistical considerations on subgroup analysis in clinical trials. *Statistics Biopharmaceutical Research, 7*(4), 286–304.

Bland, R. P., & Duncan, D. B. (1964) *On a Bayes rule for choosing the largest mean* (Paper No. 366). Biostatistics Dept., Johns Hopkins University, Baltimore.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Dixon, D. O., & Simon, R. (1991). Bayesian subset analysis. *Biometrics, 47,* 871–881.

Dixon, D. O., & Simon, R. (1992). Bayesian subset analysis in a colorectal cancer clinical trial. *Statistics in Medicine, 11,* 13–22.

Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association, 70*(350), 311–319.

Gamalo-Siebers, M., Tiwari, R., & LaVange, L. (2016). Flexible shrinkage estimation of subgroup effects through Dirichlet process priors. *Journal of Biopharmaceutical Statistics, 26*(6), 1040–1055.

Gelfand, A. E., Sahu, S. K., & Carlin, B. P. (1996). Efficient parametrizations for generalized linear mixed models. In J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith (Eds.), *Bayesian statistics 5: Proceedings of the Fifth Valencia International Meeting, June 5–9, 1994* (Oxford Science Publications) (1st ed., pp. 165–180), Oxford Science Publications, Clarendon Press, Oxford.

Gilks, W. R., & Roberts, G. O. (1996). Strategies for improving MCMC. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 89–114). New York: Chapman & Hall/CRC.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 1–19). New York: Chapman & Hall/ CRC.

Henderson, N. C., Louis, T. A., Wang, C., & Varadhan, R. (2016). Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. *Health Services and Outcomes Research Methodology, 16,* 213–233.

Holford, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics, 36*(2), 299–305.

Hsu, Y., Zalkikar, J., & Tiwari, R. C. (2017) Hierarchical Bayes approach for subgroup analysis. *Journal Biopharmaceutical Statistics.* http://journals.sagepub.com/doi/pdf/10.1177/0962280217721782. Accessed 27 August 2017.

Jewell, N. P. (2004). *Statistical for epidemiology*. New York: Chapman & Hall/CRC.

Karuri, S. W., & Simon, R. (2012). A two-stage Bayesian design for co-development of new drugs and companion diagnostics. *Statistics in Medicine, 31,* 901–914.

Kuo, L., & Smith, A. F. M. (1992). Bayesian computations in survival models via the Gibbs sampler. In J. P. Klein & P. K. Gael (Eds.), *Survival analysis: State of the art, 11–24*. The Netherlands: Kluwer Academic Publishers.

Lewis, C., & Thayer, D. T. (2004). A loss function related to the FDR for random effects multiple comparisons. *The Journal of Statistical Planning and Inference, 125,* 49–58.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10,* 325–337.

MathWorks®. Inc., (2014). MATLAB 8.4, Natick, Massachusetts, United States.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Boca Raton, FL: CRC Press.

Pennello, G. (1997). The k-ratio multiple comparisons Bayes rule for the balanced two-way design. *Journal of American Statistical Association, 92,* 675–684.

Senn, S. J. (1989). Covariate imbalance and random allocation in clinical trials. *Statististics Medicine, 8,* 467–475.

Simon, R. (2002). Bayesian subset analysis: Application to studying treatment-by-gender interactions. *Statististics Medicine, 21,* 2909–2916.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions*. New York: Springer.

US Food and Drug Administration. Clinical review, NDA 20-386/S-032, Cozaar™ (losartan potassium) tablets. https://www.fda.gov/ohrms/dockets/ac/03/briefing/3920B1_02_A-FDA-Cozaar%20Clinical%20Review.pdf (2003a). Accessed 28 August 2017.

US Food and Drug Administration. Final print label, NDA 20-386/S-032, Cozaar™ (losartan potassium) tablets. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2003/20-386S032_Cozaar_prntlbl.pdf (2003c). Accessed 28 August 2017.

US Food and Drug Administration. Statistical review, NDA 20-386/S-032, Cozaar™ (losartan potassium) tablets. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2003/20-386S032_Cozaar_statr.pdf (2003b). Accessed 28 August 2017.

Varadhan, R., & Wang, S. J. (2014). Standardization for subgroup analysis in randomized controlled trials. *Journal of Biopharmaceutical Statistics, 24*(1), 154–167.

Waller, R. A., Duncan, & D. B. (1969) A Bayes rule for the symmetric multiple comparisons problem. *Journal of the Statistical Association*, *64*, 1484–1503; Corrigenda (1972) *67*, 253–255.

Wang, C., Louis, T. A., Henderson, N. C., Weiss, C. O., & Beanz, V. R. (2018). An R package for Bayesian analysis of heterogeneous treatment effect with graphical user interface. *Journal of Statistical Software.*

Chapter 11
A Question-Based Approach to the Analysis of Safety Data

Melvin S. Munsaka

11.1 Introduction

A primary objective in the analysis of safety data is to establish a comprehensive safety profile of a drug. This is a key consideration and an area of focus in both the pre-marketing drug development and post-approval life cycle management phases. In the pre-market setting, the primary safety information comes from clinical trials data covering several domains and other supporting information, such as, safety pharmacology, toxicology, historical control data, and the literature on the therapeutic area and drug class. Reports of clinical data in the form of tables, listings, and graphs in some cases are often the main sources for assessing drug safety in the development phase. In the post-marketing setting, safety data can come from a variety of sources, including spontaneous adverse event reports, electronic health records, the literature, epidemiology studies, and more recently social media resources. Data sources for safety assessment from pre-marketing and post-marketing sources both have advantages and disadvantages that can affect generalizability of results and conclusions drawn about safety. Some regulatory guidelines are available describing expectations of how safety data are to be analyzed and reported. Often, analysis, presentation, and reporting of safety data in clinical reports tends to follow these guidelines. Whereas it is recognized that the detection of safety signals early in the drug development process is essential to minimize harms to patients and reduce late attrition due to safety issues, it is also well acknowledged that the analysis of safety data is challenging and the usual approaches may not be sufficient for a variety of reasons. In this chapter, we will discuss some considerations that pertain to the analysis and reporting of safety data. Thereafter we discuss a question-based approach to the analysis of safety data that can be used to more appropriately and

M. S. Munsaka (✉)
348 Churchill Lane, Gurnee, IL 60031, USA
e-mail: Melvin.s.munsaka@gmail.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_11

systematically look at safety data as part of the process of establishing the safety profile and informing the risk-benefit of a drug, focusing on data from clinical trials.

11.2 The Role of and a Need to Improve the Analysis and Reporting of Safety Data in Drug Development

Patient safety has always been a primary focus in drug development and there is a general consensus that patient safety must always come first. Stopping drug development for a drug candidate, drug recalls, and safety warnings in drug labels have become common with the increased scrutiny of safety data and suggests inadequacies in the characterization of the safety profiles of drugs. In fact, the adequacy of assessment of safety data had long been recognized as lacking and requiring improvement (see for example, Scherer and Wiltse 1996; Wittes 1996; Northington 1996; Tremmel 1996). This points out to an important need for a more comprehensive characterization of a drug safety profile. In this regard, there have been many efforts dedicated to finding ways of enhancing drug safety assessment and reporting in the form of new methodology and regulatory guidance. Some of the efforts on methodology for analysis of safety data, particularly those efforts directed towards quantitative methods can be seen in Jiang and Xia (2014), Gould (2015) and Gibbons and Amatya (2015). Along the same lines, regulatory and non-regulatory guidance documents have been proposed to aid in enhancement of safety analysis and reporting, for example, the FDA (2005) draft guidance for Safety Assessment for IND Safety Reporting and the Council for International Organizations of Medical Sciences Ten (CIOMS X 2017) document of meta-analysis of safety data.

The resulting and emerging recent theme is that of a need for an efficient safety analysis that facilitates for identification and characterization of the safety profile of a drug as early as possible in the development process. This includes identifying risk factors related to increased toxicity and characterization of temporal relations of adverse drug experiences and exposure and assessing magnitude of risk and its management. This heightened effort should provide information to support appropriate labeling of drugs and prevent costly consequences when a drug is marketed. In order to achieve this goal, the bar on the analysis of safety data needs to be raised by applying, or developing appropriate formal statistical methods that are helpful in identifying and characterizing safety signals in terms of various considerations such as magnitude and intensity, thereby providing a comprehensive characterization of the safety profile. In essence, there is a need to continuously monitor safety on an ongoing basis pre- and post-approval taking into account many considerations. As a matter of fact, the safety profile of a drug can evolve over time. Most importantly, the analysis of safety data should facilitate a clinician's assessment of the risk-benefit profile of the drug, classes of patients and patient management.

11.3 The Nature of Safety Data and Core Safety Data Domains

11.3.1 The Nature of Safety Data

Compared to efficacy data, safety data tend to be much more complex and highly inter-related. They are not easy to analyze with conventional statistical methods because many of the standard assumptions are not necessarily satisfied. Additionally, there are many pathological features frequently seen in safety data, including, non-normal data, high variability, and heterogeneous sub-populations. For example, patients are differentially prone to adverse events depending on their prognosis and two patients with the same prognosis can exhibit differences in their safety response and experience to treatment. Further, differences in standards of care and clinical assessment can also contribute to high variability in clinician's reporting and assessment of safety data. For example, in the assessment of severity of an adverse event, for same patient, two different clinicians can give different accounts of severity assessments and may recommend different treatment regimens for the same prognosis. A further complicating factor is that a specific adverse condition may manifest itself in different ways, or may require several pieces of related safety information to conclusively ascertain harm and causal effects.

It is worth noting that in a typical clinical trial study designed primarily to show efficacy, the majority of the data collected pertains to safety. However, based on the standard analyses of safety data in clinical trial reports, it is evident that although a lot of safety data are collected, the overall treatment of safety data is not reflective of this and it is probably and not necessarily the most appropriate data upon which to conclusively base safety decisions. In essence, the appropriate data to demonstrate the safety of a product will depend on the proposed indication, life-threatening potential, or quality of life enhancement, intended duration of use (one time versus short-term versus long-term versus intermittent versus recurrent use) and diversity of the patient population (age, race, gender, disease history, medication history, concomitant medication, concomitant disease, standard of care, genetic disposition, and many other factors).

11.3.2 Core Safety Data Domains

The core safety data domains include adverse events, clinical laboratory data, vital signs, and electrocardiograms. Other specialty safety data based on indication and class of medication may also be collected to help assess specific drug adverse effects. For example, an assessment of the eyes collected using specialty equipment may be performed to ascertain opthalmologic safety. Various adverse events of special interest, for example, the drug's effect on hypogylcemia in diabetes patients, may

require additional collection of fasting blood glucose data and systematic review for a more definitive assessment and severity classification. Below, we discuss each of the core data domains.

Adverse Events

ICH-E6 defines an adverse event as: *An adverse event (AE) is any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have a causal relationship with this treatment. An adverse event (AE) can therefore be any unfavorable and unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medicinal (investigational) product, whether or not related to the medicinal (investigational) product*. There are many considerations that need to be taken into account in the analysis and clinical assessment of adverse events, including, seriousness, severity, relatedness to drug, frequency, outcome, and expectedness. Often, analysis of adverse events revolves around magnitude of the count data (crude rates) and focuses on these aforementioned considerations with tabular output. The use of count data for adverse events as an outcome variable is complicated by the large number of possible events and incidence on placebo.

In essence, a proper analysis of adverse events requires that other considerations be addressed in order to assess degree and magnitude of potential safety signals adequately. Tremmel (1996) discussed some considerations that should be taken into account to adequately and appropriately analyze and clinically assess adverse events. These include type of event and clinical trial that is being conducted and duration of the trial. Taken together, these considerations should drive the appropriate metric, measure of risk and consequently the method of analysis. These considerations are given below in: Table 11.1 (event type), Table 11.2 (type of clinical trial and meaningful measure), and Table 11.3 (event type and analysis).

Note that in some cases, it makes more sense to look at a medical concept or adverse event of special interest (AESI) defined by a collection of adverse events that constitute or define that medical concept. This can be done through a Standardized MedDRA Query (SMQs) or a custom MedDRA Query. Some of the medical concepts may also require additional information including clinical laboratory or may involve an algorithmically-based definition. Common AESI's include: hepatotoxicity, QT interval prolongation (Torsade de Pointe), renal failure,

Table 11.1 Adverse event considerations: event type

Event type	Example	Question
Absorbing	Death	Will I get it?
Absorbing	Blindness	When will I get it?
Repeating	Seizure	How often will I get it?
Repeating	Seizure	Will I develop tolerance?
Long duration	Depressive disorders	How much time?
Long duration	Neutropenia	How much time?

Table 11.2 Adverse event considerations: type of trial and meaningful measure of risk

Type of trial	Type of AE	Meaningful measure
Short term	All	Crude rate
Short term	All	Cumulative rate
Long term	Absorbing	Hazard function
Long term	Recurring	Hazard function
Long term	Long duration	Prevalence

Table 11.3 Adverse event considerations: event type, trial duration, and analysis method

Event type	Analysis
Absorbing events	Crude incidence rate
	Events per unit time
	Survival rate (cumulative rate)
	The hazard as a function of time
Recurrent events of short duration	Events per unit time
	Expected number of events as a function of the hazard
	Hazard—simple Anderson-Gill model
	Modeling the effect of preceding events
	Heterogeneity among subjects
Recurrent events of long duration	Prevalence rates
	Markov models
	Hazard—simple Anderson-Gill model
	Modeling the effect of preceding events
	Heterogeneity among subjects

(nephrotoxicity), abuse potential, bone marrow toxicity, drug-drug interactions, polymorphic metabolism, rhabdomyolysis, pancreatitis, cardiotoxicity, hypersensitivity, serious skin reactions, non-cytotoxic bone-marrow toxicity, anaphylaxis, blindness, deafness, hemoplytic neumonia, and suicidality.

Clinical Laboratory Data

Various, laboratory data can be a manifestation of potential safety concerns. For example, increased levels of alanine aminotransferase (ALT), aspartate aminotransferase (AST), total bilirubin, and other hepatic safety parameters can be indicative of potentially compromised hepatic safety. Clinical laboratory data are essentially multivariate, non-normal, and correlated time series for a given patient. They are typically assessed on the basis of normal ranges. These limits are a univariate approach that is well known from basic multivariate distribution theory to be problematic where variables are correlated. It is quite typical to accept normal ranges at face value and can be in this sense be misinterpreted. Due to high variability in these data especially when looking across different patient populations and subgroups, it may

be more appropriate to consider using markedly abnormal values (MAVs) to identify outlying values for analysis purposes.

The usual analysis of laboratory data typically includes three standard approaches that focus on measures of central tendency (mean change from baseline), shifts from normal to abnormal, and markedly abnormal values. Per the ICH-E3 guideline, markedly abnormal values (sometimes referred to as clinically significant or marked outlier) should be defined by the sponsor. MAVs may be expressed as a proportional change above or below the reference range or as an absolute value outside the reference range that has clinical meaning. MAVs can also be defined in terms of relative change from baseline or as persistent abnormalities occurring at a specified number of visits (consecutive or not).

Vital Signs and Electrocardiogram Data

Vital signs data consisting of heart rate, blood pressure, respiration rate, temperature, height, and weight are usually collected at multiple times over the course of a clinical trial. In some cases vital signs may be collected more frequently, for example, in Phase 1 studies where there may be a need to correlate these measurements with other data such as pharmacokinetic data. Electrocardiogram (ECG) parameters are derived from ECG tracings are also increasingly collected in many studies. The key ECG parameters include: heart rate, RR, QT Interval, and QTc Intervals. ECG parameters are usually derived from traces measured at two or three time points. The derivation of these measures needs a certain skill which introduces another source of variation and interpretation. Like clinical laboratory data, both vital signs and ECG parameters data are multivariate and non-normal and like clinical laboratory data, the analysis of vital signs and ECG parameter data will typically include three standard approaches that focus on measures of central tendency (mean change from baseline), shifts from normal to abnormal, and markedly abnormal values.

Other Data Sources Important for Safety Analysis

Many other types of data domains may be collected for safety analysis purposes, including specialty safety data. Of note, there are also several other data sources that are specifically important in reporting clinical data and useful in putting safety data into some context, or for subgroup analysis, prognosis, etiology, or predictions purposes. These include demographic characteristics, drug exposure information, concomitant medications, concurrent disease, medical and medication history.

11.4 Guidance on Analysis and Inference of Safety Data

11.4.1 Guidance on Analysis of Safety Data

Suggestions for analyzing safety data are discussed in various regulatory guidance documents. For example, ICH-E9 recommends descriptive statistical methods supplemented by confidence intervals and points out that p-values are useful for

Table 11.4 Adverse event analysis method

Source	Analysis
ICH-E3	• Adverse events occurring after initiation of study treatment
	• Changes in vital signs considered as serious adverse events (SAEs)
	• Changes in laboratory parameters that were considered SAEs
	• Listing of AEs by patient
	• Listing of deaths, SAEs, and other significant AEs
FDA safety clinical review guideline	• Incidence of common AEs
	• Common AE tables
	• Identify common and drug related AEs
	• Additional analyses and explorations—age, gender, etc
CIOMS	• Rates of AEs
	• Relative risk and odd ratio
	• Confidence intervals
	• Time-to-event methods

evaluating specific differences of interest. ICH-E9 further states that if hypothesis tests are used, statistical adjustments for multiplicity to quantitate the Type I error are appropriate, but argues the Type II error is usually of more concern. It also suggests that p-values sometimes are useful as a *flagging* device applied to a large number of safety variables to highlight differences worthy of further attention. Some examples of guidance for analysis of adverse events and clinical laboratory parameters are provided in Tables 11.4 and 11.5, respectively.

Regarding the Integrated Analysis of Safety (IAS), specific guidance for integrated analyses of safety are provided in CFR 21 314.50 (d) (5) (vi) (a). Of note, it is stated that: *The applicant shall submit an integrated summary of all available information about the safety of the drug product, including pertinent animal data, demonstrated or potential adverse effects of the drug, clinically significant drug/drug interactions, and other safety considerations, such as data from epidemiological studies of related drugs. The safety data shall be presented by gender, age, and racial subgroups. When appropriate, safety data from other subgroups of the population of patients treated also shall be presented, such as for patients with renal failure or patients with different levels of severity of the disease. A description of any statistical analyses performed in analyzing safety data should also be included*.

Integrated analyses of safety are different from study level analysis due to large amounts of data. It is common to analyze these data using predefined groupings (pooling strategy) of studies with common elements. The pooling strategy is often detailed in the integrated analysis of safety statistical analysis plan. It takes into account various factors, such as the designs (e.g., double-blind versus open label), treatment, and duration of exposure, and so on. The basic idea is to pool data from rel-

Table 11.5 Clinical laboratory data analysis method

Source	Analysis
ICH-E3	• Listing of individual labs and abnormal lab values
	• Evaluation of each lab parameter
	• Laboratory values and changes from baseline over time (descriptive and categorical based abnormal values)
	• Shift tables
	• Graphs comparing initial value and on-treatment values
	• Individually clinically significant abnormalities
FDA safety clinical review guideline	• Analyses focused on measures of central tendency
	• Analyses focused on outliers or shifts from normal to abnormal
	• Marked outliers and dropouts for laboratory abnormalities
	• Additional analyses and explorations—dose dependency, time dependency, drug-demographic, drug disease and drug-drug interactions
CIOMS	• ANCOVA for lab data with baseline value as covariate with observed value or change from baseline or maximum value (most severe value)
	• Analyze binary values of lab data based on various cutoffs
	• Graphical displays—scatter plots of baseline versus post-baseline

evant/similar studies and summarize the data as if they came from one source. Pooling data can help improve the precision of incidence estimates especially for rare adverse events. It also enables assessment of trends in small subgroups of patients, such as the elderly, that may not be possible with study-level data. Outputs from pooled analyses are used to populate various sections of the common technical document (CTD), including sections 5.3.5.3 (Integrated Summary of Safety), 2.7.4 (Summary of Safety), and 2.5 (Summary of Efficacy and Safety) and the label. It is important to exercise caution when looking at analyses of safety based on pooled data and results should be cautiously interpreted as they can lead to challenges in conclusions drawn that are based on naive cumulative information. An analysis based on naively-pooled data can lead to misleading results as a consequence of Simpson's Paradox. Various suggestions have been proposed to analyze these data more appropriately, see for example, Chuang-Stein and Beltangady (2009) and Rosenkranz (2010).

Table 11.6 On the question of inferential testing

Source	Recommendation
Enas (1991)	• While not required for every AE, inferential statistical methods can be used both formally and informally to help characterize the safety profile of a new drug and help guide the resulting inferences to the broader population
FDA Guidance on conducting a safety review	• Although not strictly hypothesis testing, p-values give some feeling for the strength of the finding and should be produced for all new drug/placebo pair-wise comparisons and any p-values meeting a $p < 0.05$ level of significance should be noted
SPERT Team Crowe et al. (2009)	• For TIER 1 and TIER 2 AEs—an estimate of the risk difference, relative risk, or odds ratio is reported together with corresponding confidence intervals or p-values
ICH-E9	• Section 6.4 Statistical Evaluation—The calculation of p-values is sometimes useful, either as an aid to evaluating a specific difference of interest or as a flagging device applied to a large number of safety and tolerability variables to highlight differences worthy of further attention
Multiplicity question (Carragher 2014)	• Classical and Bayesian methods to control false positive results. An R package was developed for the methods

11.4.2 Guidance on Statistical Inference of Safety Data

The question of statistical inference in safety data is one that is often considered controversial. There are many questions and challenges and controversies and varying opinions when it comes to statistical inference of safety data. There are questions on what, when, and how to perform inference in safety data coupled with challenges in interpretation given the many statistical challenges and in general a difficult statistical testing framework. The majority of clinical studies are often powered to assess efficacy, except in those cases where the primary outcome of interest is safety, such as in the case of cardiovascular outcome studies in diabetes. There are many arguments that have been put forward, see for example, Huster (1991), against performing any form of inference when looking at safety data. On the other hand, there are many suggestions pointing out to inclusion of some form of inference for safety data. Table 11.6 is an abstraction of some of these suggestions.

In the present discussion, we take the position that statistical inference is a plausible thing to do when looking at safety data. Indeed, the use of inference appears to be the current norm in the analysis of safety data. However, caution should be exercised in the findings from such inference and must be balanced with clinical implications, discernement, and plausibility.

11.5 A Tiered Approach to the Analysis of Safety Data

The 3-tier system (see for example, Crowe et al. 2009) for analyzing safety data is based on the premise that it is important to report all adverse events. However, not all adverse events need to be analyzed in the same manner. Based on this three tier system, adverse events are classified into 1 of 3 tiers for analysis purposes. A key important feature of this approach is to distinguish between prespecified hypotheses which fall into the Tier 1 category and events that are not prespecified, the Tier 2 and Tier 3 categories. The key consideration is that is prespecified hypotheses should be handled differently from those not prespecified. The specification of the three tiers should be included in each trial protocol and the prospective safety analysis plan. Note that serious adverse events are included in each tier. The definitions of the 3 tiers of AEs are as follows:

Tier 1 events are prespecified for detailed analysis and hypothesis testing. These are events for which a pre-specified hypothesis has been defined. In general, multiplicity adjustment are not to be used for Tier 1 events, but this can be considered if there are numerous Tier 1 events. Of note, appropriate metrics of absolute risk should be reported, such, frequency, subject incidence, or incidence rate per person-time of exposure. Estimates of the risk difference, relative risk, or odds ratio can be reported together with corresponding confidence intervals or p-values. Additionally, risk factors such as age, sex, and co-morbidities can be investigated as predictors of these events within each treatment group and overall.

Tier 2 are targeted for signal detection among common events. This tier of adverse events are those that do not have a pre-specified hypothesis and are common. The Tier 2 events should be reported with risk differences, risk ratios, or odds ratios including confidence intervals and/or p-values. The events in Tier 2 are the events for which signal detection and multiplicity adjustment, if necessary, should be considered.

The Tier 3 events are those events that are not in the Tier 1 or Tier 2 designation. These events are reported with descriptive statistics, typically, number and percent (n, %) and possibly rates per person-time, but without p-values or confidence intervals. It is important to note that inclusion in Tier 3 does not imply lack of importance, particularly for clinically serious events.

The three tier approach to the analysis of safety data provides a general framework on how one might categorize the types of adverse events for analysis purposes. However, complex safety issues may need more than a single, or even multiple, Tier 1 event to be fully understood. As noted earlier, comprehensive assessment of certain medical concepts may need to be looked into using SMQs or customized MedDRA queries and draw upon other evidentiary data. For example, to perform a comprehensive assessment of liver safety, one needs to look at not only individual adverse events. Rather, one should consider the SMQ for hepatic events and also look at supporting evidence from laboratory abnormalities associated with hepatic safety, and make use of various analyses such as Hy's law in combination with AE data, and various other considerations such as those outlined in the FDA (2009) guidance for drug induced liver injury (DILI) or the DILIN network criteria (Aithal et al. 2011).

11.6 Challenges in Reporting and Analysis of Safety Data

11.6.1 Reporting Safety Data

There are many challenges in the analysis and reporting of safety data and some of these have already been highlighted in earlier sections and are discussed widely in the literature, see for example Singh and Loke (2012). They include lack of an evidentiary gold standard, limited statistical power, lack of adequate ascertainment and classification of adverse events, and limited generalizability. Also, as noted in the earlier sections, clinical trials collect a great deal of data relating to the safety of the trial participants. The data are complex in nature and traditional approaches to data review involve using summary tables and listed data. Safety data also present many challenges with regard to analysis and interpretation. The very nature of safety data makes it challenging to analyze using conventional statistical methods because many standard assumptions may not be fulfilled.

Additionally, a typical clinical trial is generally not sufficient to detect safety signals, unless a study is specifically powered for safety. The pathological features of diseases lead to asymmetric nonnormal distributions and heterogeneous subpopulations. Often descriptive tabular outputs with lots of exploration and review of individual patient data are primary source for safety assessment. This use of tabular outputs for safety data often results in large volumes of output leading to problems in generation, assessment, validation, assembly and last and worst of all interpretation, comprehension, and communication of key safety findings, leading to challenges in the overall interpretation and decision making. The simple descriptive summary tables and review of individual patient data are rarely analytical with lots of exploration and estimation. As pointed out by Wittes (1996): *A plethora of tables and graphs that describe safety may bury some true signal in a cacophony of numbers. The simple descriptive summary tabular outputs and the review of individual patient data are rarely analytical*. Rarely is there comprehensive analytical approaches and inference to better ascertain the safety profile of the drug which can aid in decision making.

Reporting of AEs in randomized clinical trials (RCTs) is often lacking and with limited application in the real world, as RCTs are of short duration, include small numbers of patients, and are selective for subjects lacking in comorbid conditions. It is not surprising that new and unexpected safety concerns emerge with any new drug after it has been launched and used by many more patients. Part of the problem is inherent to the way safety data are reported in RCTs. The typical clinical trial is generally not sufficient to detect safety signals, unless study is specifically powered for safety—zero observed events does not mean drug is safe. The pathological features leading to asymmetric non-normal distributions, heterogeneous sub-populations, high variability in measurements, and multi-dimensional and inter-related in nature of the safety data make it difficult to analyze. The key safety endpoints of concern may not be known prior to trial.

It is also difficult to design a clinical trial to simultaneously provide for complete and comprehensive inference about all safety effects and efficacy of a drug. In fact, for the most part, clinical trial entry criteria and designs are targeted at efficacy assessment. There are also challenges in reporting and in general, typical study-level clinical trial data are generally not sufficient to conclusively assess safety. Even for those studies that are specifically geared for safety, the focus is often on a specific adverse event of special interest. In any event, some adverse events of interest may not be known a priori. Often, safety analysis tends to be somewhat ad hoc and exploratory in nature. Interestingly, there is often more safety related data collected than efficacy data in terms of volume, but even with this large quantity of safety data acquired during clinical drug testing, safety data are rarely harvested to their fullest potential.

11.6.2 Analysis of Safey Data

It is evident that the characterization of the safety profile of a drug requires analyses, both descriptive and inferential, that go beyond the usual common tabular presentation of safety data. In general, these additional analyses may be targeted analyses geared towards addressing one or more specific safety considerations. It is also widely acknowledged that there is some room for improvement in the analysis and reporting of safety data from clinical trials and that safety data needs to be given a more rigorous treatment similar to efficacy. In fact, much is written in old and recent literature about the inadequacies and incompleteness of the statistical evaluation and reporting of clinical safety data. This evolution in safety data analysis needs and reporting has resulted in a shift in the roles within regulatory science and sponsor companies, with both parties needing to allocate more resources to look into safety data in a more systematic way in an attempt to appropriately provide a comprehensive assessment of the safety profile of a drug.

The analysis of safety data typically hinges on individual AEs often looking at these in isolation of other AEs, pointing out various notable differences between treatment where deemed appropriate. AE analysis focuses on how many subjects experience the AE in question and in some cases the total number of reports of the AE in question. This analysis approach to AEs neglects the fact that AEs are not independent of each other. Specifically, this analysis approach ignores the potential concurrence of AEs within patients as well as other information such as the number of occurrences and time-course of the AEs within patients. Further, information about the concurrence or constellation of AEs within patients is a valuable piece of information but this is usually neglected. The information regarding AEs that occur together or in some constellation can have potential applications in patient treatment and care. For the most part, results from both efficacy and safety analyses eventually end up in the label in one form or another, but really do not speak much to the data in terms of the issues mentioned above, including patient management and care.

Standard/typical analysis of clinical trial data often follows the usual approach of separately analyzing efficacy and safety data which does not really render itself useful when assessing patient management and risk-benefit. The predominant method for statistical evaluation and interpretation of safety data collected in a clinical trial is the tabular display of descriptive statistics. There is a great opportunity to enhance evaluation of drug safety through, for example, the use of graphical displays, which can convey multiple pieces of information concisely and more effectively than can tables.

Many suggestions have been put forward to use alternative quantitative methods as an alternative to the common tabular outputs for exploring safety data and that these methods present a great opportunity to enhance evaluation of drug safety. To this end quantitative statistical evaluation of safety data is evolving into a major component of the totality of evaluation of drug product. The analysis of safety data from clinical trials offers unique methodological opportunities. Some common approaches for analyzing adverse events are described in Cao and He (2011), including crude percentage (rate) and adverse events adjusted by exposure time or recurrent. Overall, the analyses of safety data are not often rigorously done and often fail to account for a variety of characteristics that are pertinent to safety data. The general consensus is that there is room for improvement in the analysis and reporting of safety data from clinical trials and that safety data are often not adequately and appropriately assessed with more focus given to efficacy and with often selective reporting for safety data. There is insufficient use of more appropriate methodological approaches for safety data with more rigor given to analysis of efficacy data despite the current atmosphere of more scrutiny on safety, both pre- and post-approval. As a consequence, various methods have been proposed in the literature aimed at addressing some of the safety considerations. To this end, we highlight below three examples of some suggestions to improve analysis of safety data.

Harrell (2005) pointed out that it is difficult to see patterns in tables and substituting graphs for tables can help increase efficiency of review. Graphs can be used to aid in inference and communicating safety results and to help display large amounts of safety data coherently and maximize the ability to detect unusual features or patterns. They can also play a big role in facilitating communication of safety results with regulators, investigators, Data Monitoring Committees, and other stakeholders. Visualization of safety data can help convey multiple pieces of information concisely and more effectively than tables. Graphical exploration can substantially improve information gain from safety data.

The usual standard analysis of adverse events using crude rates is known to be problematic. For example, the required statistical assumptions of constant hazard rate over time for valid estimates of incidence rates are not likely to be met in practice by adverse events data of clinical trials. In this setting, a non-parametric approach called the mean cumulative function can provide a valid statistical inference on recurrent adverse event profiles of drugs in randomized clinical trials, see for example, Cao

and He (2011) and Siddiqui (2009). Wang and Quartey (2012) also proposed use of a nonparametric method to estimate the mean cumulative duration based on the nonparametric cumulative mean function estimate.

In practice, the usual inferential approach for safety data analysis involves the comparison of the proportion of subjects who experience an AE between treatment groups for each type of AE. This involves a large number of analyses with inadequate statistical power and no meaningful control of type 1 error. Unadjusted analyses can lead to false positive results, while using simple adjustments, for example, the Bonferroni adjustment, are generally too conservative and counterproductive for considerations of safety. Thus, an important consideration in the analysis of AE data is to address concerns about multiplicity and the imprecision of data inherent in the AE data. Analyses of AE data are routinely analyzed using p-values. If p-values are reported and interpreted without multiplicity considerations, there can be incorrect conclusions due to false positive findings which can needlessly muddy the safety profile of an otherwise safe drug. Various approaches for dealing with the multiplicity question have been proposed within the safety data setting. Although there is no harmonized agreement on using one method over the other, they provide a reasonable balance between no adjustment versus adjustment. Several frequentist and Bayesian methods have been proposed to address the multiplicity issue. For an overview of these methods, see for example, Carragher (2015) who also developed an R package that implements these methods (https://cran.r-project.org/web/packages/c212/index.html).

11.7 A Question-Based Approach to the Analysis of Safety Data

11.7.1 A Question-Based Approach

As noted in the previous sections, analysis and reporting of safety data is much more complex than efficacy data and wrought with many challenges. Various suggestions have been proposed to tackle those challenges and there is a reasonable level of mathematical sophistication and processes that can be put into place to facilitate better analysis of safety data. The degree of detail, sophistication of analysis, and language used should in general be determined by target audience which can be, for example, regulatory authorities or pharmacovilance personnel evaluating the safety of the drug.

Additionally, it is clear that every analysis that is performed for safety data can clearly be associated with a specific safety concern or concerns. It thus makes sense to systematically tackle the analysis of safety data taking all these into account and

employing various planning, analysis, and reporting suggestions that have been put forward such as those pertaining to the use of the tiered approach and multiplicity adjustments. We suggest the use of a question-based approach in analyzing safety data. The approach can be used in conjunction with the various proposals made to help improve the analysis of safety data. The premise is that each safety concern can best be addressed by asking appropriate questions and then identifying the data that can be used to address the question followed by identifying a methodological approach (descriptive, graphical, analytical or inferential) to address the question and ultimately reporting and decision-making associated with the findings.

It is worth to note that asking questions about safety data is not a new idea. Many authors have discussed questions that need to be addressed and also proposed some ways to address those questions when looking at safety data. For example, Durham and Turner (2008) discuss a set of patient-centric questions as part of the rationale for evaluating safety data in clinical trials. For example, one question that they ask is: *how likely is that my patient will experience an adverse event reaction that is so serious that it may be life threatening?* Merz et al. (2014) also discuss a set of questions that they consider in their paper on methodology to assess clinical liver safety data. For example, one question that they ask is: *Are there any Hy's law cases in the dataset?* Harrell (2005) also considered this approach in addressing safety issues. Among the questions he discussed are: *Who is having the selected AEs? Which AEs occur together? Which AEs tend to occur in the same patient?* Within the context of subgroup analysis, Chow and Liu (2009) also posed the following questions among others: *Are the AE rates the same across a subgroup for patient taking the drug? Within subgroup levels, are AE rates the same across treatment groups? Is the time of occurrence of the AE the same across levels of a subgroups?*

The specific approach that we are proposing here is that the analysis of safety data can be done more appropriately and systematically by asking a set of question surrounding a safety issue of concern. These questions will then dictate the data that can be used to address the question at hand and consequently the method to be used (descriptive, graphical, analytical, or inferential). By employing various analysis approaches that have been proposed in the literature to address safety concerns, this can best be done within a question-based setting. In fact any given analysis method of safety data can be associated with a specific question that it is trying to address and hence the data that is needed and conclusions and decisions that will be drawn regarding the question being asked. The following graphic in Fig. 11.1 summarized this approach.

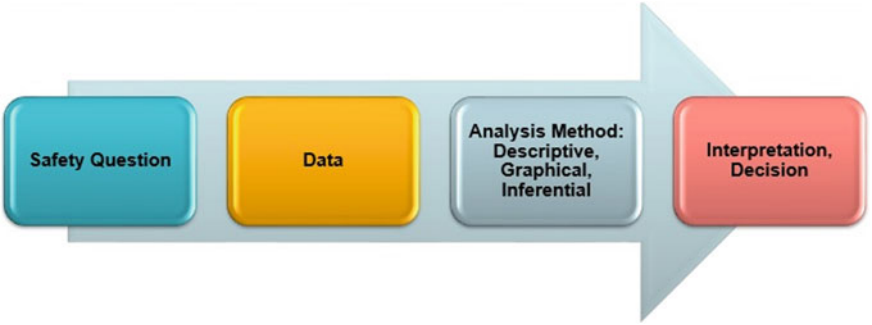


Fig. 11.1 Clinical laboratory data analysis method

11.7.2 Sample Questions

For illustration purposes, below we provide some questions that one may ask for adverse events and separately for clinical laboratory data and subgroups.

Questions on Adverse Events

- What is the temporal relation of drug experience and exposure?
- Which AEs are elevated in treatment versus control?
- Is there any evidence of a dose-response relationship?
- Is there a difference in the time to the first event across treatment groups?
- What are the AE durations?
- What are the trends of time to the first event among different AEs?
- What is the severity of the AEs?
- Which AEs tend to occur in the same patient?
- Are there withdraws and/or interruption due to AE of interest?
- Is there a relationship with other AEs?
- Which AEs are occurring together in clusters or in a constellation?
- Is the potential AE of interest increasing over time?
- What are the risk factors of the AE?
- Is there a relationship with use of concomitant medications?
- Are the most prevalent AEs suggestive of more serious events or medical concern?

Questions on Clinical Laboratory Data

- How many patients exceed certain threshold values of clinical laboratory data across treatment groups?
- Are distributions of clinical laboratory data and incidence of out-of-range values different across treatment groups?
- Are there subjects with elevation on multiple clinical laboratory data?
- What is the distribution and magnitude of clinical laboratory data elevations?
- Is there any evidence of a dose-response relationship in clinical laboratory data elevations?
- What is the timing of clinical laboratory data abnormalities?
- Is there a characteristic time to event for clinical laboratory data elevations?
- Are shifts from baseline different between treatment groups for some clinical laboratory data?
- Is there any evidence for a dose-response-relationship for clinical laboratory data exceeding thresholds?
- What is the time-course of clinical laboratory data elevations?
- What are the patterns of elevations, e.g., single versus multiple/recurrent?
- What is the duration of elevations?
- Are clinical laboratory data elevations transient?
- Do the clinical laboratory data elevations resolve on treatment or off-treatment?
- Can we identify potential risk factors or subgroups of patients associated with clinical laboratory data elevations?

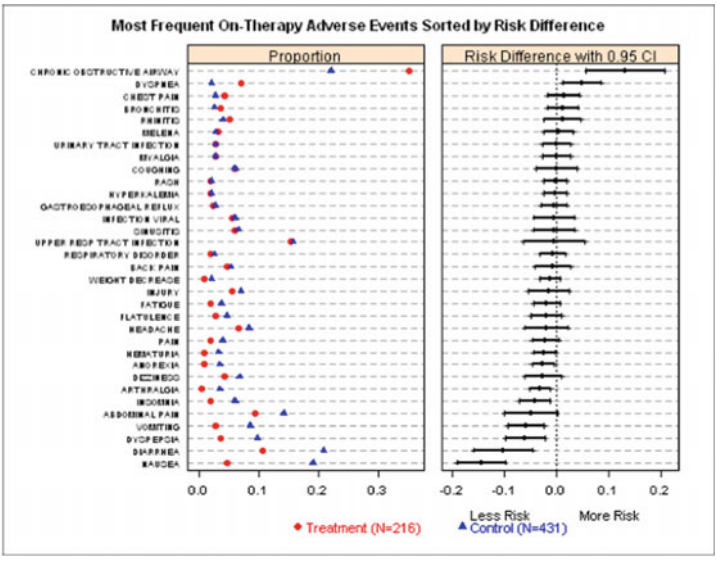
Questions on Subgroups

- Are adverse event rates the same across a subgroup for patients taking experimental drugs?
- In which subgroups do the AEs occur?
- Within subgroup levels, are AEs the same across treatment groups?
- Is there a consistent association between the treatment group and the adverse event response across levels of a subgroup?
- Does the subgroup predict an adverse event response?
- Is the time the occurrence of the adverse event the same across levels of a subgroup?

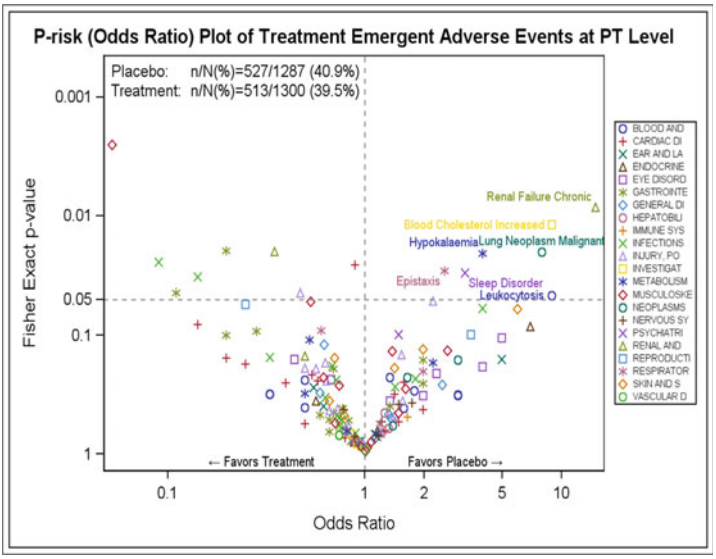
11.7.3 Examples

For illustration purposes, we consider some examples of the question-based approach by drawing on some graphs from the website: http://www.ctspedia.org/do/view/CTSpedia/StatGraphHome. Specifically, we consider how each graphic can be used to address a safety question of interest.

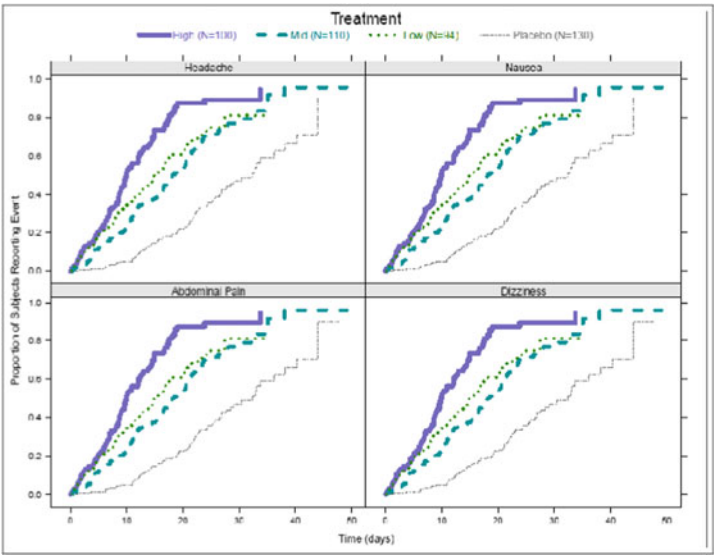
Question: *What are the frequencies, magnitudes, and differences between treatment and control in some pre-specified adverse events?* Note that in this particular setting, we are looking a some pre-specified events, possibly Tier 1 events. One approach to address this question is to use a risk plot, such as the one shown below which would provide answers to the question being asked. An additional detail here is the provision of confidence intervals for the risk difference.



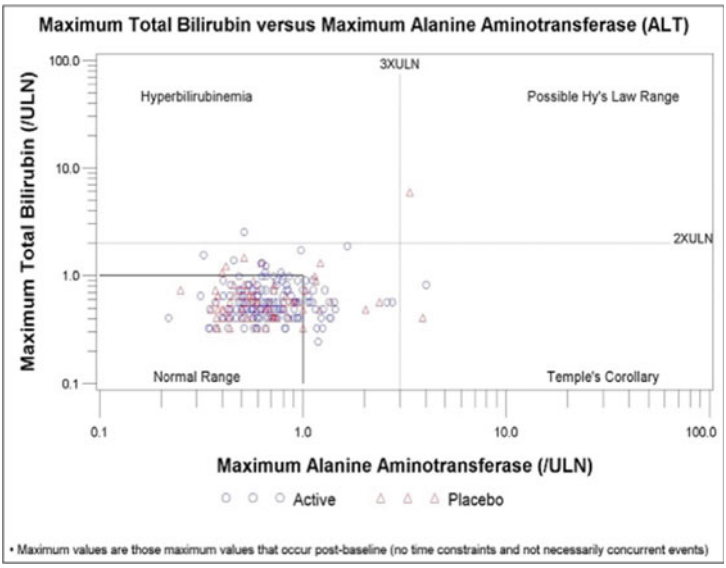
Question: *Are there any other events that we should be concerned about?* Note that in this particular setting, we are not pre-specifying the events. Hence one can envisage that we are looking at Tier 2 events. One approach to address this question is to use a volcano plot, such as the one shown below which would provide answers to the question being asked. The general idea is to focus on the AEs whose p-value exceeds a particular threshold, in this case $p = 0.05$ and odds ratios exceeding 1. One would then focus on the AEs in the upper right quadrant. The thresholds can be modified as deemed appropriate and the risk difference and relative risk can be used in place of the odds ratio. Further, this can be used in conjuction with methods to accounting for multiplicity.



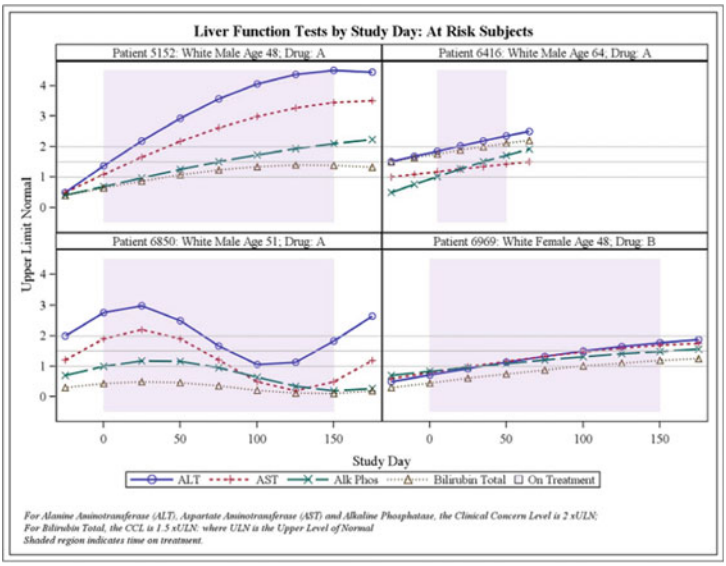
Question: *What is the timing of headache, dizziness, nausea, and vomiting?* One approach to address this question is to simply plot the cumulative incidence of each adverse event and visually inspect the plots to see if there are differences in the pattern of timing of the events. Other types of plots can also be used to address this question, including the Kaplan-Meier plot, hazard-plot, and event charts, with each type of plot having different interpretation. Once can also use the same the figure to plot all the events for ease of comparison.



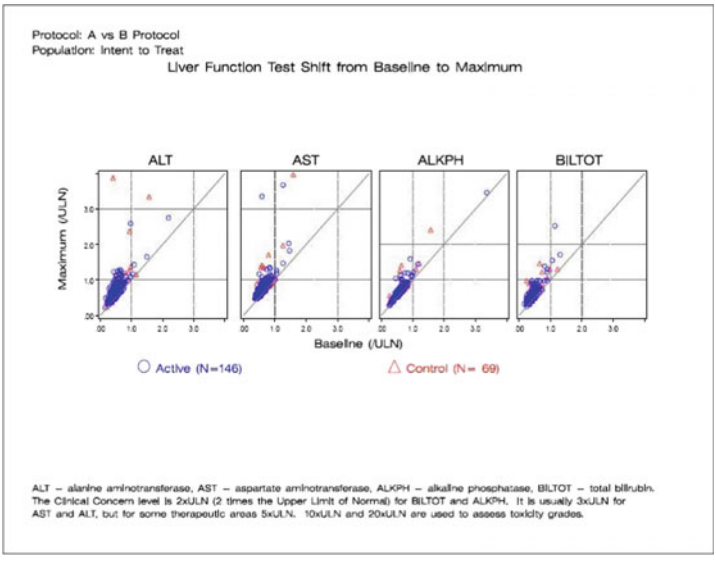
Question: *Are there subjects meeting Hy's law?* The most common approach to address this question is to use the eDISH (evaluation of Drug-Induced Serious Hepatotoxicity) plot. One form of the e-DISH plot is shown below. Basically one focuses on the upper quadrant to see if any subjects fall in that quadrant, known as Hy's Law Range, which is indicative of a potential liver safety problem. This can then be used with other information to conclusively rule out other causes besides the drug. This often includes looking at the individual patient data that include the time course of elevations of liver safety-related clinical laboratory parameters and how these parameters track together. Examples of individual patient figures of time course of clinical laboratory parameters for liver safety clinical laboratory and a figure how to assess how these track together are given after the eDISH plot, respectively.



Question: *What is the time course of liver safety laboratory parameters for selected patients?*



Question: *Which liver safety laboratory parameters have some elevations?*



All the graphs presented above to address different questions can be accompanied or supplemented by appropriate and/or additional analytical methods to further elucidate the safety issue of concern. Each graph can also address more than one question.

11.8 Conclusion

This chapter discussed several considerations associated with the analysis of safety focusing on the pre-market setting. It was noted that the analysis of safety data is challenging and the usual approaches may not be sufficient for a variety of reasons. It was also noted that some regulatory guidelines are available describing expectations of how safety data are to be analyzed and reported and often, analysis, presentation, and reporting of safety data in practical settings tends to follow these guidelines. Various considerations and challenges that pertain to the analysis and reporting of safety data were highlighted. It was also noted that it is well recognized that current standard approaches to the analysis of safety data are lacking in various aspects and there is room for improvement. In particular, it was noted that safety analyses should not only provide rapid answers to pre-specified questions, but also insight into the structure of raw data and also generate new questions and provide information on safety profile of the drug, at a minimum point out what risks are associated with the drug. It should also facilitate for more efficient identification of potential signals early in development process, convey safety information more efficiently, identify trends and patterns in potential adverse events of interest. It would make safety results more understandable quantitatively, increase likelihood of detecting key safety signals, improve ability to make clinical decisions, help in decision making regarding specific safety concerns, provide basis for systematic exploration safety concerns.

In this regard, there have been many concerted efforts dedicated to finding better ways of enhancing drug safety assessment and reporting in the form of new methodology, particularly those efforts directed towards quantitative methods, and regulatory guidance. We discussed a question-based approach to the analysis of safety data that can be used to more appropriately and systematically look at safety data as part of the process of establishing the safety profile of a drug. Examples based on graphical approaches were provided to help illustrate the question-based approach. It was argued that the question-based approach used in conjunction with the emerging quantitative methodology for safety data can play a big role for an efficient safety analysis. This in turn can facilitate for identification and characterization of the safety profile of a drug as early as possible in the development process and result in less attrition of drugs due to safety issues and enable a well thought-out process for safety analysis. Most importantly, this approach can also help in facilitating for clinicians assessment of the risk-benefit profile of the drug and classes of patients and patient management.

References

Aithal, G. P., et al. (2011). Case definition and phenotype standardization in drug-induced liver injury. *Clinical Pharmacology and Therapeutics*, *89*, 806–815.

Berry, S. M., & Berry, D. A. (2004). Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics*, *60*, 418–426.

Cao, D., & He, X. (2011). Statistical analysis of adverse events in randomized clinical trials using SAS. PharmaSUG2011, Paper SP07. http://www.lexjansen.com/pharmasug/2011/SP/PharmaSUG-2011-SP07.pdf.

Carragher, R. (2013). A comparison of some methods for detection of safety signals in randomised controlled clinical trials. http://www.sctweb.org/public/meetings/2015/slides/CPS%2013%20-%20Carragher.pdf.

Carragher, R. (2015). A comparison of some methods for detection of safety signals in randomised controlled trials. https://strathprints.strath.ac.uk/57234/1/Carragher_SCT_2015_methods_for_detection_of_safety_signals_in_randomised_controlled_trials.pdf.

Chow, S-C., & Liu, J-P. (2009). *Design and analysis of clinical trials: Concepts and methodologies*. New York: Wiley.

Chuang-Stein, C., & Beltangady, M. (2009). Reporting cumulative proportion of subjects with an adverse event based on data from multiple studies. *Pharmaceutical Statistics*, *10*, 3–7.

CIOMS X (2017). *Evidence synthesis and meta-analysis for drug safety*. http://cioms.ch/shop/product/evidence-synthesis-and-meta-analysis-report-of-cioms-working-group-x/.

Chuang-Stein, C., & Xia, H. A. (2013). The practice of pre-marketing safety assessment in drug development. *Journal of Biopharmaceutical Statistics*, *23*, 13–25.

Crowe, B. J., et al. (2009). Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: A report of the safety planning, evaluation, and reporting team. *Clinical Trials*, *6*, 430–440.

Durham, T. A., & Turner, J. R. (2008). *Introduction to statistics in pharmaceutical clinical trials*. London: Pharmaceutical Press.

Enas, G. (1991). Making decisions about safety in clinical trials—The case for inferential statistics. *Drug Information Journal*, *25*, 439–446.

FDA (2005). Drug clinical trials need prospective safety analysis plans. *The Pink Sheet, 67*(1).

FDA (2009). Guidance for industry drug-induced liver injury: Premarketing clinical evaluation. https://www.fda.gov/downloads/Drugs/guidances/UCM174090.pdf.

FDA (2015). Safety Assessment for IND Safety Reporting. https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM477584.pdf.

Gibbons, R. D., & Amatya, A. (2015). *Statistical methods for drug safety*. Boca Raton: Chapman and Hall/CRC.

Gould, A. L. (2015). *Statistical methods for evaluating safety in medical product development*. Statistics in Practice. New York: Wiley.

Harrell (2005). *Exploratory analysis of clinical safety data to detect safety signals*. http://biostat.mc.vanderbilt.edu/wiki/pub/Main/FHHandouts/gsksafety.pdf.

Herson, J. (2016). *Data and safety monitoring committees in clinical trials*. Boca Raton: Chapman and Hall/CRC.

Huster, W. J. (1991). Clinical trial adverse events: The case for descriptive techniques. *Drug information journal, 25*, 447–456.

ICH-E3: Guideline for Industry Structure and Content of Clinical Study Reports. https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM073113.pdf.

ICH-E6: Guideline for Good Clinical Practice. https://www.fda.gov/downloads/Drugs/Guidances/ucm073122.pdf.

ICH-E9: Statistical Principles for Clinical Trials. https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf.

Jiang, Q., & Xia, H. A. (2014). *Quantitative evaluation of safety in drug development: Design, analysis and reporting*. Boca Raton: Chapman and Hall/CRC.

Jiang, Q., & He, W. (2016). *Benefit-risk assessment methods in medical product development: Bridging qualitative and quantitative assessments*. Boca Raton: Chapman and Hall/CRC.

Lakshminarayanan, M., & Kaur, A. (2008). Safety findings in clinical findings: Are they real or just coincidental. In *Proceeings ASA Biopharmaceutical Section*, pp. 2203–2209.

Liu, J.-P. (2007). Rethinking statistical approaches to evaluating drug safety. *Yonsei Medical Journal*, *48*, 895–900.

Ma, H., Ke, C., Jiang, Q., & Snappinn, S. (2015). Statistical considerations on the evaluation of imbalances of adverse events in randomized clinical trials. *Therapeutic Innovation and Regulatory Science*, *49*, 957–965.

Mehrotra, D. V., & Heyse, J. F. (2004). Use of the false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research*, *13*, 227–38.

Merz, M., Lee, K. R., Kullak-Ublick, G. A., Brueckner, A., & Watkins, P. B. (2014). Methodology to assess clinical liver safety data. *Drug Safety*, *37*(Suppl 1), S33–S45.

Nilsson, M. E., & Koke, S. C. (2001). Defining treatment-emergent adverse events with the medical dictionary for regulatory activities (MedDRA). *Drug Information Journal*, *15*, 1289–1299.

Northington, B. (1996). A review of issues in the collection and reporting of adverse events. *Biopharmaceutical Report*, *4*, 1–5.

O'Neill, R. T. (1987). Statistical analysis of adverse event data from clinical trials: Special emphasis on serious events. *Drug Information Journal*, *21*, 9–20.

Rosenkranz, G. K. (2010). An approach to integrated safety analyses from clinical studies. *Drug Information Journal*, *44*, 649–657.

Scherer, J. C., & Wiltse, C. G. (1996). Adverse events: After 58 years, do we have it right yet? *Biopharmaceutical Report*, *4*, 1–5.

Siddiqui, O. (2009). Statistical methods to analyze adverse events data of randomized clinical trials. *Journal of Biopharmaceutical Statistics*, *19*, 889–999.

Singh, S., & Loke, Y. K. (2012). Drug safety assessment in clinical trials: Methodological challenges and opportunities. *Trials*, *13*(138). http://www.trialsjournal.com/content/13/1/138.

Temple, E. J. (1991). The regulatory eveolution of the intergrated safety summary. *Drug Information Journal*, *15*, 1289–1299.

Tremmel, L. (1996). Describing risk in long term clinical trials. *Biopharmaceutical Report*, *4*, 5–8.

Wang, J., & Quartey, G. (2012). Nonparametric estimation for cumulative duration of adverse events. *Biometrical Journal*, *54*, 61–74.

Wittes, J. (1996). A statistical perspective on adverse event reporting in clinical trials. *Biopharmaceutical Report*, *4*, 5–10.

Xia, H. A., & Jiang, Q. (2014). Statistical evaluation of drug safety data. *Therapeutic Innovation and Regulatory Science*, *48*, 109–120.

Xia, H. A., Ma, H., & Carlin, B. P. (2011). Bayesian hierarchical modeling for detecting safety signals in clinical trials. *Journal of Biopharmaceutical Statistics*, *21*, 1006–1029.

Xia, H. A., Crowe, B. J., Schriver, R. C., Oster, M., & Hall, D. B. (2011). Planning and core analyses for periodic aggregate safety data reviews. *Clinical Trials*, *8*, 175–182.

Chapter 12
Analysis of Two-Stage Adaptive Trial Designs

Shein-Chung Chow and Min Lin

12.1 Introduction

In the past decade, adaptive design methods in clinical research have attracted much attention because it offers not only the principal investigators potential flexibility for identifying clinical benefit of a test treatment under investigation, but also efficiency for speeding up the development process. The FDA adaptive design draft guidance defines an adaptive design as a clinical study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study (FDA 2010). As it is recognized by many investigators/researchers, the use of adaptive design methods in clinical trials may allow the researchers to correct assumptions used at the planning stage and select the most promising option early. In addition, adaptive designs make use of cumulative information of the on-going trial, which provide the investigator an opportunity to react earlier to surprises regardless of positive or negative results Thus, the adaptive design approaches may speed up the drug development process.

Despite the possible benefits for having a second chance to modify the trial at interim when utilizing an adaptive design, it can be more problematic operationally due to bias that may have introduced to the conduct of the trial. As indicated by the FDA draft guidance, operational biases may occur when adaptations in trial and/or statistical procedures are applied after the review of interim (unblinded) data. As a result, it is a concern whether scientific integrity and validity of trial are warranted. Chow and Chang (2011) indicated that trial procedures include, but not limited

S.-C. Chow (✉)
Duke University School of Medicine, Durham, NC, USA
e-mail: sheinchung.chow@duke.edu

M. Lin (✉)
Food and Drug Administration, Silver Spring, MD, USA
e-mail: min.lin@fda.hhs.gov

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_12

to, inclusion/exclusion criteria, dose/dose regimen and treatment duration, endpoint selection and assessment, and/or laboratory testing procedures employed. On the other hand, statistical procedures are referred to as study design, statistical hypotheses (which can reflect study objectives), endpoint selection, power analysis for sample size calculation, sample size re-estimation, and/or sample size adjustment, randomization schedules, and statistical analysis plan (SAP). With respect to these trial and statistical procedures, commonly employed adaptations at interim include, but are not limited to, (1) sample size re-estimation at interim analysis, (2) adaptive randomization with unequal treatment allocation (e.g., change from 1:1 ratio to 2:1 ratio), (3) deleting, adding, or modifying treatment arms after the review of interim data, (4) shifting in patient population due to protocol amendment, (5) different statistical methods, (6) changing study endpoints (e.g., change response rate and/or survival to time-to-disease progression in cancer trials), and (7) changing hypotheses/objectives (e.g., switch a superiority hypothesis to a non-inferiority hypothesis). Therefore, the use of the adaptive design methods in clinical trials seems promising because of its potential flexibility for identifying any possible clinical benefit, signal, and/or trend regarding efficacy and safety of the test treatment under investigation. However, major adaptations may have an impact on the integrity and validity of the clinical trials, which may raise some critical concerns to the accurate and reliable evaluation of the test treatment under investigation. These concerns include (1) that the control of the overall type I error rate at a pre-specified level of significance, (2) that the correctness of the obtained *p*-values, and (3) that the reliability of the obtained confidence interval. Most importantly, major (significant) adaptations may have resulted in a totally different trial that is unable to address the scientific/medical questions the original study intended to answer.

Chow (2011) indicated that a seamless trial design is defined as a trial design that combines two independent trials into a single study that can addresses study objectives from individual studies. An adaptive seamless design is referred to as a seamless trial design that would use data collected before and after the adaptation in the final analysis. In practice, a two-stage seamless adaptive design typically consists of two stages (phases): a learning (or exploratory) phase (Stage 1) and a confirmatory phase (Stage 2). The objective of the learning phase is not only to obtain information regarding the uncertainty of the test treatment under investigation but also to provide the investigator the opportunity to stop the trial early due to safety and/or futility/efficacy based on accrued data or to apply some adaptations such as adaptive randomization at the end of Stage 1. The objective of the second stage is to confirm the findings observed from the first stage. A two-stage seamless adaptive trial design has the following advantages that (1) it may reduce lead time between studies (the traditional approach); (2) it provides the investigator the second chance to re-design the trial after the review of accumulated date at the end of Stage 1. Most importantly, data collected from both stages are combined for a final analysis in order to fully utilize all data collected from the trial for a more accurate and reliable assessment of the test treatment under investigation.

12.2 Types of Two-Stage Adaptive Designs

Chow and Tu (2008) and Chow (2011) classified two-stage seamless adaptive trial designs into the following four categories depending upon study objectives and study endpoints at different stage.

Table 12.1 indicates that there are four different types of two-stage adaptive trial designs depending upon whether study objectives and/or study endpoints at different stages are the same. For example, Category I designs (i.e., SS designs) include those designs with same study objectives and same study endpoints, while Category II and Category III designs (i.e., SD and DS designs) are referred to those designs with same study objectives but different study endpoints and different study objectives but same study endpoints, respectively. Category IV designs (i.e., DD designs) are the study designs with different study objectives and different study endpoints. In practice, different study objectives could be treatment selection for Stage 1 and efficacy confirmation for Stage 2. On the other hand, different study endpoints could be biomarker, surrogate endpoints, or a clinical endpoint with a shorter duration at the first stage versus clinical endpoint at the second stage. Note that a group sequential design with one planned interim analysis is often considered an SS design.

In practice, typical examples for a two-stage adaptive seamless design include a two-stage adaptive seamless phase I/II design and a two-stage adaptive seamless phase II/III design. For the two-stage adaptive seamless phase I/II design, the objective at the first stage may be for biomarker development and the study objective for the second stage is usually to establish early efficacy. For a two-stage adaptive seamless phase II/III design, the study objective is often for treatment selection (or dose finding) while the study objective at the second stage is for efficacy confirmation. In this article, our focus will be placed on Category II designs. The results can be similarly applied to Category III and Category IV designs.

It should be noted that the terms seamless and phase II/III were not used in the FDA draft guidance as they have sometimes been adopted to describe various design features (FDA 2010). In this article, a two-stage adaptive seamless phase II/III design only refers to a study containing an exploratory phase II stage (Stage 1) and a confirmatory phase III stage (Stage 2) while data collected at both phases (stages) will be used for final analysis.

One of the questions that are commonly asked when applying a two-stage adaptive seamless design in clinical trials is sample size calculation/allocation. For the first kind (i.e. Category I, SS) of two-stage seamless designs, the methods based on

Table 12.1 Types of two-stage adaptive designs

	Study	Endpoint
Study objectives	Same (S)	Different (D)
Same (S)	I = SS	II = SD
Different (D)	III = DS	IV = DD

Source Chow (2011)

individual *p*-values as described in Chow and Chang (2011) can be applied. However, for other kinds (i.e. Category II to Category IV) of two-stage seamless trial designs, standard statistical methods for group sequential design are not appropriate and hence should not be applied directly. For Category II–IV trial designs, power analysis and/or statistical methods for data analysis are challenging to the biostatistician. For example, a commonly asked question is "How do we control the overall type I error rate at a pre-specified level of significance?" In the interest of stopping trial early, "How to determine stopping boundaries?" is a challenge to the investigator and the biostatistician. In practice, it is often of interest to determine whether the typical O'Brien-Fleming type of boundaries is feasible. Another challenge is "How to perform a valid analysis that combines data collected from different stages?" To address these questions, Cheng and Chow (2016) proposed the concept of a multiple-stage transitional seamless adaptive design which takes into consideration of different study objectives and study endpoints.

12.3 Analysis SS Two-Stage Adaptive Designs

Category I design with same study objectives and same study endpoints at different stages is considered similar to a typical group sequential design with one planned interim analysis. Thus, standard statistical methods for group sequential design are often employed. It, however, should be noted that with various adaptations that applied, these standard statistical methods may not be appropriate. In practice, many interesting methods for Category I designs are available in the literature. These methods include (1) Fisher's criterion for combining independent *p*-values (Bauer and Kohne 1994; Bauer and Rohmel 1995; Posch and Bauer 2000), (2) weighted test statistics (Cui et al. 1999), (3) the conditional error function approach (Liu and Chi 2001; Proschan and Hunsberger 1995), and (4) conditional power approaches (Li et al. 2005).

Among these methods, Fisher's method for combining *p*-values provides great flexibility in selecting statistical tests for individual hypotheses based on sub-samples. Fisher's method, however, lacks flexibility in the choice of boundaries (Muller and Schafer 2001). For Category I adaptive designs, many related issues have been studied. For example, Rosenberger and Lachin (2003) explored the potential use of response-adaptive randomization. Chow et al. (2005) examined the impact of population shift due to protocol amendments. Li et al. (2005) studied a two-stage adaptive design with a survival endpoint, while Hommel et al. (2005) studied a two-stage adaptive design with correlated data. An adaptive design with a bivariate-endpoint was studied by Todd (2003). Tsiatis and Mehta (2003) showed that there exists a more powerful group sequential design for any adaptive design with sample size adjustment.

For illustration purpose, in what follows, we will introduce the method based on sum of *p*-values (MSP) by Chang (2007) and Chow and Chang (2011). The MSP

follows the idea of considering a linear combination of the p-values from different stages.

12.3.1 Theoretical Framework for Multiple-Stage Adaptive Designs

Consider a clinical trial utilizing a K-stage design. This is similar to a clinical trial with K interim analyses, while the final analysis is the Kth interim (final) analysis. Suppose that at each interim analysis, a hypothesis test is performed. The objective of the trial can be formulated as the following intersection of the individual hypothesis tests from the interim analyses

$$H_0 : H_{01} \cap \cdots \cap H_{0K},$$

where $H_{0i},\ i = 1, \ldots, K$ is the null hypothesis to be tested at the ith interim analysis. Note that there are some restrictions on H_{0i}, that is, rejection of any $H_{0i},\ i = 1, \ldots, K$ will lead to the same clinical implication (e.g. drug is efficacious); hence all $H_{0i},\ i = 1, \ldots, K$ are constructed for testing the same endpoint within a trial. Otherwise the global hypothesis cannot be interpreted.

In practice, H_{0i} is tested based on a sub-sample from each stage, and without loss of generality, assume H_{0i} is a test for the efficacy of a test treatment under investigation, which can be written as

$$H_{0i} : \eta_{i1} \geq \eta_{i2} \quad \text{versus} \quad H_{ai} : \eta_{i1} < \eta_{i2},$$

where η_{i1} and η_{i2} are the responses of the two treatment groups at the ith stage and we assume bigger values are better. It is often the case that when $\eta_{i1} = \eta_{i2}$, the p-value p_i for the sub-sample at the ith stage is uniformly distributed on [0, 1] under H_0. Under the null hypothesis, Bauer and Kohne (1994) used Fisher's combination of the p-values to construct a test statistic for multiple-stage adaptive designs. Following similar idea, Chang (2007) and Chow and Chang (2011) considered a linear combination of the p-values as follows

$$T_k = \sum_{i=1}^{K} w_{ki}\, p_i,\ i = 1, \ldots, K, \tag{12.1}$$

where $w_{ki} > 0$ and K is the number of interim analyses planned. If $w_{ki} = 1$, this leads to

$$T_k = \sum_{i=1}^{K} p_i,\ i = 1, \ldots, K. \tag{12.2}$$

T_k can be viewed as cumulative evidence against H_0. Thus, the smaller the T_k is, the stronger the evidence is. Alternatively, we can consider $T_k = \sum_{i=1}^{K} p_i/K$, which is an average of the evidence against H_0. Intuitively, one may consider the stopping rules

$$\begin{cases} \text{Stop for efficacy} & \text{if } T_k \le \alpha_k \\ \text{Stop for futility} & \text{if } T_k \ge \beta_k \\ \text{Continue} & \text{otherwise} \end{cases}, \tag{12.3}$$

where T_k, α_k, and β_k are monotonic increasing functions of k, $\alpha_k < \beta_k$, $k = 1, \ldots, K-1$, and $\alpha_K = \beta_K$. Note that α_k and β_k are referred to as the efficacy and futility boundaries, respectively. To reach the kth stage, a trial has to pass 1 to $(k-1)$th stages. Therefore, a so-called proceeding probability can be defined as the following unconditional probability:

$$\begin{aligned} \psi_k(t) &= P(T_k < t, \alpha_1 < T_1 < \beta_1, \ldots, \alpha_{k-1} < T_{k-1} < \beta_{k-1}) \\ &= \int_{\alpha_1}^{\beta_1} \cdots \int_{\alpha_{k-1}}^{\beta_{k-1}} \int_{-\infty}^{t} f_{T_1 \cdots T_k}(t_1, \ldots, t_k) dt_k dt_{k-1} \cdots dt_1, \end{aligned} \tag{12.4}$$

where $t \ge 0$, t_i, $i = 1, \ldots, k$ is the test statistic at the ith stage, and $f_{T_1 \cdots T_k}$ is the joint probability density function. Thus, the error rate at the kth stage can be obtained as

$$\pi_k = \psi_k(\alpha_k). \tag{12.5}$$

Since the type I error rates at different stages are mutually exclusive, the experiment-wise type I error rate is sum of $\pi_k, k = 1, \ldots, K$. Thus, we have

$$\alpha = \sum_{k=1}^{K} \pi_k. \tag{12.6}$$

Note that stopping boundaries can be determined with appropriate choices of α_k. The adjusted p-value calculation is the same as the one in a classic group sequential design (Jennison and Turnbull 2000). The key idea is that when the test statistic at the kth stage $T_k = t = \alpha_k$ (i.e. just on the efficacy stopping boundary), the p-value is equal to alpha spent $\sum_{i=1}^{k} \pi_i$. This is true regardless of which error spending function is used and consistent with the p-value definition of the traditional design. As indicated in Chang (2007), the adjusted p-value corresponding to an observed test statistic $T_k = t$ at the kth stage can be defined as

$$p(t; k) = \sum_{i=1}^{k-1} \pi_i + \psi_k(t), \ k = 1, .., K. \tag{12.7}$$

Note that p_i in Eq. (12.1) is the stage-wise (unadjusted) p-value from a sub-sample at the ith stage, while $p(t;k)$ are adjusted p-values calculated from the test statistic, which are based on the cumulative sample up to the kth stage where the trial stops, Eqs. (12.6) and (12.7) are valid regardless how p_i are calculated.

12.3.2 Two-Stage Design

In this section, for simplicity, we will consider the method of sum of p-values (MSP) and apply the general framework to the two-stage designs as outlined in Chang (2007) and Chow and Chang (2011) which are suitable for the following adaptive designs that allow (1) early efficacy stopping, (2) early stopping for both efficacy and futility; and (3) early futility stopping. These adaptive designs are briefly described below.

Early efficacy stopping—For simplicity, consider $K = 2$ (i.e., a two-stage design) which allows for early efficacy stopping (i.e., $\beta_1 = 1$). By (12.5), the type I error rates to spend at Stage 1 and Stage 2 are given by

$$\pi_1 = \psi_1(\alpha_1) = \int_0^{\alpha_1} dt_1 = \alpha_1, \tag{12.8}$$

and

$$\pi_2 = \psi_2(\alpha_2) = \int_{\alpha_1}^{\alpha_2} \int_t^{\alpha_1} dt_2 dt_1 = \frac{1}{2}(\alpha_2 - \alpha_1)^2, \tag{12.9}$$

respectively. Using Eqs. (12.8) and (12.9), (12.6) becomes

$$\alpha = \alpha_1 + \frac{1}{2}(\alpha_2 - \alpha_1)^2. \tag{12.10}$$

Solving for α_2, we obtain

$$\alpha_2 = \sqrt{2(\alpha - \alpha_1)} + \alpha_1. \tag{12.11}$$

α_1 is the stopping probability (error spent) at the first stage under the null hypothesis condition and $\alpha - \alpha_1$ is the error spent at the second stage. As a result, if the test statistic $t_1 = p_1 > \alpha_2$, it is certain that $t_2 = p_1 + p_2 > \alpha_2$. Therefore, the trial should stop when $p_1 > \alpha_2$ for futility.

Based on relationship among $\alpha_1, \alpha_2,$ and α as given in (12.10), various stopping boundaries can be considered with appropriate choices of $\alpha_1, \alpha_2,$ and α. For illustration purpose, Table 12.2 provides some examples of the stopping boundaries from Eqs. (12.10) and (12.11).

By (12.7)–(12.11), the adjusted p-value is given by

Table 12.2 Stopping boundaries for two-stage efficacy designs

One-sided α	α_1	0.005	0.010	0.015	0.020	0.025	0.030
0.025	α_2	0.2050	0.1832	0.1564	0.1200	0.0250	–
0.05	α_2	0.3050	0.2928	0.2796	0.2649	0.2486	0.2300

Source Chang (2007). Statistics in Medicine, 26, 2772–2784

Table 12.3 Stopping boundaries for two-stage efficacy and futility designs

One-sided α	$\beta_1 = 0.15$					
0.025	α_1	0.005	0.010	0.015	0.020	0.025
	α_1	0.2154	0.1871	0.1566	0.1200	0.0250
	$\beta_1 = 0.20$					
0.05	α_2	0.005	0.010	0.015	0.020	0.025
	α_2	0.3333	0.3155	0.2967	0.2767	0.2554

Source Chang (2007). Statistics in Medicine, 26, 2772–2784

$$p(t;k) = \begin{cases} t & \text{if } k = 1 \\ \alpha_1 + \frac{1}{2}(t-\alpha_1)^2 & \text{if } k = 2 \end{cases}, \tag{12.12}$$

where $t = p_1$ if the trial stops at Stage 1 and $t = p_1 + p_2$ if the trial stops at Stage 2.

Early efficacy or futility stopping—For this case, it is obvious that if $\beta_1 \geq \alpha_2$, the stopping boundary is the same as it is for the design with early efficacy stopping. However, futility boundary β_1 when $\beta_1 \geq \alpha_2$ is expected to affect the power of the hypothesis testing. Therefore,

$$\pi_1 = \int_0^{\alpha_1} dt_1 = \alpha_1, \tag{12.13}$$

and

$$\pi_2 = \begin{cases} \int_{\alpha_1}^{\beta_1} \int_{t_1}^{\alpha_2} dt_2 dt_1 & \text{for } \beta_1 \leq \alpha_2 \\ \int_{\alpha_1}^{\alpha_2} \int_{t_1}^{\alpha_2} dt_2 dt_1 & \text{for } \beta_1 > \alpha_2 \end{cases} \tag{12.14}$$

Thus, it can be verified that

$$\alpha = \begin{cases} \alpha_1 + \alpha_2(\beta_1 - \alpha_1) - \frac{1}{2}(\beta_1^2 - \alpha_1^2) & \text{for } \beta_1 < \alpha_2 \\ \alpha_1 + \frac{1}{2}(\alpha_2 - \alpha_1)^2 & \text{for } \beta_1 \geq \alpha_2 \end{cases} \tag{12.15}$$

Similarly, under (12.15), various boundaries can be obtained with appropriate choices of $\alpha_1, \alpha_2, \beta_1,$ and α (Table 12.3). The adjusted p-value is given by

Table 12.4 Stopping boundaries for two-stage futility design

One-sided α	β_1	0.1	0.2	0.3	≥0.4
0.025	α_2	0.3000	0.2250	0.2236	0.2236
0.05	α_2	0.5500	0.3500	0.3167	0.3162

Source Chang (2007). Statistics in Medicine, 26, 2772–2784

$$p(t;k) = \begin{cases} t & \text{if } k = 1 \\ \alpha_1 + t(\beta_1 - \alpha_1) - \frac{1}{2}(\beta_1^2 - \alpha_1^2) & \text{if } k = 2 \text{ and } \beta_1 < \alpha_2 \\ \alpha_1 + \frac{1}{2}(t - \alpha_1)^2 & \text{if } k = 2\ \beta_1 \geq \alpha_2 \end{cases} \tag{12.16}$$

where $t = p_1$ if the trial stops at Stage 1 and $t = p_1 + p_2$ if the trial stops at Stage 2.

For a trial design with early futility stopping, it is a special case of the previous design, where $\alpha_1 = 0$ in Eq. (12.15). Hence, we have

$$\alpha = \begin{cases} \alpha_2\beta_1 - \frac{1}{2}\beta_1^2 & \text{for } \beta_1 < \alpha_2 \\ \frac{1}{2}\alpha_2^2 & \text{for } \beta_1 \geq \alpha_2 \end{cases} \tag{12.17}$$

Solving for α_2, we have

$$\alpha_2 = \begin{cases} \frac{\alpha}{\beta_1} + \frac{1}{2}\beta_1 & \text{for } \beta_1 < \sqrt{2\alpha} \\ \sqrt{2\alpha} & \text{for } \beta_1 \geq \alpha_2 \end{cases} \tag{12.18}$$

Table 12.4 gives examples of the stopping boundaries generated using Eq. (12.18). The adjusted p-value can be obtained from Eq. (12.16), where $\alpha_1 = 0$, that is,

$$p(t;k) = \begin{cases} t & \text{if } k = 1 \\ \alpha_1 + t\beta_1 - \frac{1}{2}\beta_1^2 & \text{if } k = 2 \text{ and } \beta_1 < \alpha_2 \\ \alpha_1 + \frac{1}{2}t^2 & \text{if } k = 2\ \beta_1 \geq \alpha_2 \end{cases} \tag{12.19}$$

12.3.3 Conditional Power

Conditional power with or without clinical trial simulation is often considered for sample size re-estimation in adaptive trial designs. As discussed earlier, since the stopping boundaries for the most existing methods are either based on z-scale or p-value, to link a z-scale and a p-value, we will consider $p_k = 1 - \Phi(z_k)$ or inversely, $z_k = \Phi^{-1}(1 - p_k)$, where z_k and p_k are the normal z-score and the p-value from the sub-sample at the kth stage, respectively. It should be noted that z_2 has asymptotically

normal distribution with $N(\delta/se(\hat{\delta}_2), 1)$ under the alternative hypothesis, where $\hat{\delta}_2$ is the estimation of treatment difference in the second stage and

$$se(\hat{\delta}_2) = \sqrt{2\hat{\sigma}^2/n_2} \approx \sqrt{2\sigma^2/n_2}.$$

The conditional power can be evaluated under the alternative hypothesis when rejecting the null hypothesis H_0. That is,

$$z_2 \geq B(\alpha_2, p_1). \tag{12.20}$$

Thus, the conditional probability given the first stage naïve p-value, p_1 at the second stage is given by

$$P_C(p_1, \delta) = 1 - \Phi\left(B(\alpha_2, p_1) - \frac{\delta}{\sigma}\sqrt{\frac{n_2}{2}}\right), \ \alpha_1 < p_1 \leq \beta_1. \tag{12.21}$$

As an example, for the method based on the product of stage-wise p-values (MPP), the rejection criterion for the second stage is

$$p_1 p_2 \leq \alpha_2, \ \text{i.e.,} \ z_2 \geq \Phi^{-1}(1 - \alpha_2/p_1).$$

Therefore, $B(\alpha_2, p_1) = \Phi^{-1}(1 - \alpha_2/p_1)$.

Similarly, for the method based on the sum of stage-wise p-values (MSP), the rejection criterion for the second stage is

$$p_1 + p_2 \leq \alpha_2, \ \text{i.e.,} \ z_2 = B(\alpha_2, p_1) = \Phi^{-1}(1 - \max(0, \alpha_2 - p_1)).$$

On the other hand, for the inverse normal method (Lehmacher and Wassmer 1999), the rejection criterion for the second stage is

$$w_1 z_1 + w_2 z_2 \geq \Phi^{-1}(1 - \alpha_2),$$

That is, $z_2 \geq (\Phi^{-1}(1 - \alpha_2) - w_1 \Phi^{-1}(1 - p_1))/w_2$, where w_1 and w_2 are prefixed weights satisfying the condition of $w_1^2 + w_2^2 = 1$. Note that the group sequential design and CHW method (Cui et al. 1999) are special cases of the inverse-normal method. Since the inverse normal method requires two additional parameters (w_1 and w_2), for simplicity, we will only compare the conditional powers of MPP and MSP. For a valid comparison, the same α_1 is used for both methods. As it can be seen from Eq. (12.21), the comparison of the conditional power is equivalent to the comparison of function $B(\alpha_2, p_1)$. Equating the two $B(\alpha_2, p_1)$, we have

$$\frac{\hat{\alpha}_2}{p_1} = \tilde{\alpha}_2 - p_1, \tag{12.22}$$

where $\hat{\alpha}_2$ and $\tilde{\alpha}_2$ are the final rejection boundaries for MPP and MSP, respectively. Solving (12.22) for p_1, we obtain the critical point for p_1

$$\eta = \frac{\tilde{\alpha}_2 \mp \sqrt{\tilde{\alpha}_2^2 - 4\tilde{\alpha}_2}}{2}. \tag{12.23}$$

Equation (12.23) indicates that when $p_1 < \eta_1$ or $p_2 > \eta_2$, MPP has a higher conditional power than that of MSP. When $\eta_1 < p_1 < \eta_2$, MSP has a higher conditional power than MPP.

Note that the unconditional power P_w is nothing but the expectation of conditional power, i.e.

$$P_w = E_\delta[P_C(p_1, \delta)]. \tag{12.24}$$

Therefore, the difference in unconditional power between MSP and MPP is dependent on the distribution of p_1, and consequently, dependent on the true difference δ, and the stopping boundaries at the first stage (α_1, β_1).

Note that in Bauer and Kohne's method using Fisher's combination (Bauer and Kohne 1994), which leads to the equation $\alpha_1 + \ln(\beta_1/\alpha_1)e^{-(1/2)\chi^2_{4,1-\alpha}} = \alpha$, it is obvious that determination of β_1 leads to a unique α_1, consequently α_2. This is a non-flexible approach. However, it can be verified that the method can be generalized to $\alpha_1 + \alpha_2 \ln \beta_1/\alpha_1 = \alpha$, where α_2 does not have to be $e^{-(1/2)\chi^2_{4,1-\alpha}}$.

12.4 Analysis SD Two-Stage Adaptive Designs

For illustration purpose, consider a two-stage phase II/III seamless adaptive designs which have same study objectives but different study endpoints. In what follows, we will consider the cases of continuous, binary responses, and time-to-event endpoints, respectively.

12.4.1 Continuous Endpoints

Let x_i be the observed value of the study endpoint (e.g., a biomarker) from the ith subject in phase II (Stage 1), $i = 1, \ldots, n$ and y_j be the observed value of the study endpoint (i.e. the primary clinical endpoint) from the jth subject in phase III (Stage 2), $j = 1, \ldots, m$. Suppose that $x_i's$ and $y_j's$ are independently and identically distributed with $E(x_i) = \nu$ and $Var(x_i) = \tau^2$, and $E(y_j) = \mu$ and $Var(y_j) = \sigma^2$, respectively. Chow et al. (2007) proposed obtaining predicted values of the clinical endpoint based on data collected from the biomarker (or surrogate endpoint) under an established relationship between the biomarker and the clinical endpoint. These pre-

dicted values are then be combined with the data collected at the confirmatory phase (Stage 2) to derive a statistical inference on the treatment effect under investigation. For simplicity, suppose that x and y can be correlated in the following straight-line relationship

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{12.25}$$

where ε is the random error with zero mean and variance ς^2. ε is assumed to be independent of x. In practice, we assume that this relationship is well-established. In other words, the parameters β_0 and β_1 are assumed known. Based on Eq. (12.25), the observations x_i observed in the first stage can then be transformed $\beta_0 + \beta_1 x_i$ (denoted by $\hat{y}_i$). $\hat{y}_i$ is then considered as the observation of the clinical endpoint and combined with those observations y_i collected in the second stage to estimate the treatment mean μ. Chow et al. (2007) proposed the following weighted-mean estimator,

$$\hat{\mu} = \omega\bar{\hat{y}} + (1-\omega)\bar{y} \tag{12.26}$$

where $\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$, $\bar{y} = \frac{1}{m}\sum_{j=1}^{m} y_j$ and $0 \le \omega \le 1$. It should be noted that $\hat{\mu}$ is the minimum variance unbiased estimator among all weighted-mean estimators when the weight is given by

$$\omega = \frac{n/(\beta_1^2\tau^2)}{n/(\beta_1^2\tau^2) + m/\sigma^2} \tag{12.27}$$

if β_1, τ^2 and σ^2 are known. In practice, τ^2 and σ^2 are usually unknown and ω is commonly estimated by

$$\hat{\omega} = \frac{n/s_1^2}{n/s_1^2 + m/s_2^2} \tag{12.28}$$

where s_1^2 and s_2^2 are the sample variances of $\hat{y}_i$'s and y_j's, respectively. The corresponding estimator of μ, which is denoted by

$$\hat{\mu}_{GD} = \hat{\omega}\bar{\hat{y}} + (1-\hat{\omega})\bar{y}, \tag{12.29}$$

and is referred to as the Graybill-Deal (GD) estimator of μ (Graybill and Deal 1959). Note that Meier (1953) proposed an approximate unbiased estimator of the variance of the GD estimator, which has bias of order $O(n^{-2} + m^{-2})$. Khatri and Shah (1974) gave an exact expression of the variance of this estimator in the form of an infinite series, which is given as.

$$\widehat{Var}\left(\hat{\mu}_{GD}\right) = \frac{1}{n/S_1^2 + m/S_2^2}\left[1 + 4\hat{\omega}\left(1-\hat{\omega}\right)\left(\frac{1}{n-1} + \frac{1}{m-1}\right)\right].$$

Based on the GD estimator, the comparison of the two treatments can be made by testing the following hypotheses

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_a : \mu_1 \neq \mu_2. \tag{12.30}$$

Let $\hat{y}_{ij}$ be the predicted value (based on $\beta_0 + \beta_1 x_{ij}$), which is used as the prediction of y for the *j*th subject under the *i*th treatment in phase II (Stage 1). From Eq. (18.29), the GD estimator of μ_i is given by

$$\hat{\mu}_{GDi} = \hat{\omega}_i \bar{\hat{y}}_i + (1 - \hat{\omega}_i)\bar{y}_i, \tag{12.31}$$

where $\bar{\hat{y}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \hat{y}_{ij}$, $\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$ and $\hat{\omega}_i = \frac{n_i/S_{1i}^2}{n_i/S_{1i}^2 + m_i/S_{2i}^2}$ with S_{1i}^2 and S_{2i}^2 being the sample variances of $(\hat{y}_{i1}, \ldots, \hat{y}_{in_i})$ and $(y_{i1}, \ldots, y_{im_i})$, respectively. For hypotheses (12.30), consider the following test statistic

$$\widetilde{T}_1 = \frac{\hat{\mu}_{GD1} - \hat{\mu}_{GD2}}{\sqrt{\widehat{Var}\left(\hat{\mu}_{GD1}\right) + \widehat{Var}\left(\hat{\mu}_{GD2}\right)}}, \tag{12.32}$$

where

$$\widehat{Var}\left(\hat{\mu}_{GDi}\right) = \frac{1}{n_i/S_{1i}^2 + m_i/S_{2i}^2}\left[1 + 4\hat{\omega}_i\left(1 - \hat{\omega}_i\right)\left(\frac{1}{n_i - 1} + \frac{1}{m_i - 1}\right)\right]$$

is an estimator of $Var(\hat{\mu}_{GDi})$, i = 1, 2. Consequently, an approximate $100(1 - \alpha)\%$ confidence interval of $\mu_1 - \mu_2$ is given as

$$\left(\hat{\mu}_{GD1} - \hat{\mu}_{GD2} - z_{\alpha/2}\sqrt{V_T},\ \hat{\mu}_{GD1} - \hat{\mu}_{GD2} + z_{\alpha/2}\sqrt{V_T}\right) \tag{12.33}$$

where $V_T = Var(\hat{\mu}_{GD1}) + Var(\hat{\mu}_{GD2})$. As a result, the null hypothesis H_0 is rejected if the above confidence interval does not contain 0.

12.4.2 Binary Responses

Consider the case where the primary study endpoint is a binary response with different treatment durations at different stages. Suppose that the study duration of the first[t] stage is L, while the study duration of the second stage is CL with C > 1. Assume that the response is determined by the lifetime t, and the corresponding lifetime distribution for the test treatment is $G_1(t, \theta_1)$, while for the control is $G_2(t, \theta_2)$. Denote by r_i the number of responders among n_i individuals in the *i*th stage for the test treatment, $i = 1, 2$. Similarly, denote by s_i the number of responders among m_i individuals in the *i*th stage for the control treatment. Based on the observed

data, suppose $G_1(t, \theta_1) = G(t, \lambda_1)$ and $G_2(t, \theta_2) = G(t, \lambda_2)$. Then the likelihood functions become

$$L(\lambda_i) = \left(1 - e^{-\lambda_i cL}\right)^{r_i} e^{-(n_i - r_i)\lambda_i cL} \left(1 - e^{-\lambda_i L}\right)^{s_i} e^{-(m_i - s_i)\lambda_i L} \tag{12.34}$$

Let $\hat{\lambda}_i$ be the maximum likelihood estimate (MLE) of λ_i. Utilizing numerical methods such as Newton-Raphson method, $\hat{\lambda}_i$ can be found by the solving the following equation

$$\frac{r_i c}{e^{\lambda_i cL} - 1} + \frac{s_i}{e^{\lambda_i L} - 1} - (n_i - r_i)c - (m_i - s_i) = 0, \tag{12.35}$$

which is obtained by setting the first order partial derivative with respect to the parameter to zero. Note that the MLE of λ_i exist only and only if r_i/n_i and s_i/m_i does not equal 0 or 1 at the same time.

Based on asymptotic normality of MLE, $\hat{\lambda}_i$ asymptotically follows a normal distribution. In particular, as n_i and m_i tend to infinity, $(\hat{\lambda}_i - \lambda_i)/\sigma_i(\lambda_i)$ follows the standard normal distribution where

$$\sigma_i(\lambda_i) = L^{-1}\left(n_i c^2 (e^{\lambda_i cL} - 1)^{-1} + m_i (e^{\lambda_i L} - 1)^{-1}\right)^{-1/2}.$$

Let $\sigma_i(\hat{\lambda}_i)$ be the MLE of $\sigma_i(\lambda_i)$. Then based on the consistency of MLE, by the Slutsky's Theorem $(\hat{\lambda}_i - \lambda_i)/\sigma_i(\hat{\lambda}_i)$ asymptotically follows the standard normal distribution. Consequently, an approximated $(1-\alpha)$ confidence interval of λ_i is given as $\left(\hat{\lambda}_i - z_{\alpha/2}\sigma_i(\hat{\lambda}_i),\ \hat{\lambda}_i + z_{\alpha/2}\sigma_i(\hat{\lambda}_i)\right)$, where z_u is the upper u-quantile of the standard normal distribution. Under the exponential model, comparison of two treatments usually focuses on the hazard rate λ_i. As a result, hypotheses testing for different types of comparison can be derived.

Test for equality—For equality testing, the hypotheses are formulated as

$$H_0 : \lambda_1 = \lambda_2 \quad \text{versus} \quad H_1 : \lambda_1 \neq \lambda_2 \tag{12.36}$$

Since $\hat{\lambda}_i$ is asymptotically normal distributed, and $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are independent, it follows that under the null hypothesis, $(\hat{\lambda}_1 - \hat{\lambda}_2)/\sqrt{\sigma_1^2(\hat{\lambda}_1) + \sigma_2^2(\hat{\lambda}_2)}$ asymptotically follows the standard normal distribution. Thus, the null hypothesis in (12.36) is rejected at approximate α level of significance if

$$\left|\hat{\lambda}_1 - \hat{\lambda}_2\right| \Big/ \sqrt{\sigma_1^2(\hat{\lambda}_1) + \sigma_2^2(\hat{\lambda}_2)} > z_{\alpha/2}.$$

Test for Superiority—Under the exponential model, a smaller hazard rate indicates a better performance of the treatment. As a result, to identify superiority of the new treatment over the control, the following hypotheses are considered.

$$H_0 : \lambda_2 - \lambda_1 \leq \delta \quad \text{versus} \quad H_1 : \lambda_2 - \lambda_1 > \delta, \tag{12.37}$$

where $\delta > 0$ is a difference of clinical importance. Obviously, the null hypothesis should be rejected for large value of $(\hat{\lambda}_2 - \hat{\lambda}_1 - \delta)/\sqrt{\sigma_1^2(\hat{\lambda}_1) + \sigma_2^2(\hat{\lambda}_2)}$. Under the null hypothesis, $(\hat{\lambda}_2 - \hat{\lambda}_1 - \delta)/\sqrt{\sigma_1^2(\hat{\lambda}_1) + \sigma_2^2(\hat{\lambda}_2)}$ is asymptotically normal distributed. Thus, the null hypothesis is rejected at approximately α level of significance if

$$(\hat{\lambda}_2 - \hat{\lambda}_1 - \delta)/\sqrt{\sigma_1^2(\hat{\lambda}_1) + \sigma_2^2(\hat{\lambda}_2)} > z_\alpha.$$

Test for Non-inferiority—To show that the new treatment is not worse than the control, we may consider the following hypotheses $H_0 : \lambda_1 - \lambda_2 \geq \delta$ versus $H_1 : \lambda_1 - \lambda_2 < \delta$, which are equivalent to

$$H_0 : \lambda_2 - \lambda_1 \leq -\delta \quad \text{versus} \quad H_1 : \lambda_2 - \lambda_1 > -\delta, \tag{12.38}$$

where $\delta > 0$ is a difference of clinical importance. The hypotheses in (12.38) are of similar form as those for superiority testing. Therefore, the null hypothesis is rejected at approximate α level of significance if

$$(\hat{\lambda}_2 - \hat{\lambda}_1 + \delta)/\sqrt{\sigma_1^2(\hat{\lambda}_1) + \sigma_2^2(\hat{\lambda}_2)} > z_\alpha,$$

Test for Equivalence—In clinical trial, it is commonly unknown whether the performance of new treatment is better than the (active) control, especially when prior knowledge of the new treatment is not available. In this case, it is more appropriate to consider the following hypotheses for therapeutic equivalence:

$$H_0 : |\lambda_1 - \lambda_2| \geq \delta \quad \text{versus} \quad H_1 : |\lambda_1 - \lambda_2| < \delta. \tag{12.39}$$

The above hypotheses can be tested by constructing the confidence interval of $\lambda_2 - \lambda_1$. It can be verified that the null hypothesis is rejected at a significance level α if and only if the $100(1 - 2\alpha)\%$ confidence interval $\hat{\lambda}_2 - \hat{\lambda}_1 \pm z_\alpha\sqrt{\sigma_1^2(\hat{\lambda}_1) + \sigma_2^2(\hat{\lambda}_2)}$ falls within $(-\delta, \delta)$. In other words, the test treatment is concluded to be equivalent to the control if

$$(\hat{\lambda}_2 - \hat{\lambda}_1 - \delta)/\sqrt{\sigma_1^2(\hat{\lambda}_1) + \sigma_2^2(\hat{\lambda}_2)} < -z_\alpha,$$

and

$$(\hat{\lambda}_2 - \hat{\lambda}_1 + \delta)/\sqrt{\sigma_1^2(\hat{\lambda}_1) + \sigma_2^2(\hat{\lambda}_2)} > z_\alpha.$$

12.4.3 Time-to-Event Endpoints

For illustration purpose, we will consider a two-stage adaptive clinical trial design for comparing two treatment groups, i.e., a test (T) treatment and a control or reference (R) treatment under Cox's proportional hazards model. The case under Weibull distribution can be similarly derived (see, e.g., Lu et al. 2012).

Let n_j be total sample size for the two treatments in the jth stage, $j = 1, 2$ and d_j be number of distinct failure times in the jth stage, which are denoted by $t_{j1} < t_{j2} < \cdots < t_{jd_j}$. Furthermore, denote the observation based on the kth subject in the jth stage by

$$\left(T_{jk}, \delta_{jk}, z_{jk}(t),\ 0 \le t \le T_{jk}\right) = \left(\min(\widetilde{T}_{jk}, C_{jk}),\ I(\widetilde{T}_{jk} < C_{jk}),\ z_{jk}(t),\ 0 \le t \le T_{jk}\right),$$

where, correspondingly, T_{jk} is the observed time, $\widetilde{T}_{jk}$ is time-to-event, δ_{jk} is the indicator for the observed failure, C_{jk} is a censoring time which is assumed to be independent of $\widetilde{T}_{jk}$, and $z_{jk}(t)$ is a covariate vector at time t. Let $h(t|z)$ be the hazard rate at time t for an individual with a covariate vector z. The Cox proportional hazard model (Cox 1972) assumes

$$h(t|z(t)) = h(t|0)e^{b'z(t)},$$

where the baseline $h(t|0)$ is unspecified and b is a coefficient vector with the same dimension as $z(t)$. Thus, the partial likelihood function is

$$\begin{aligned} L(b) &= \prod_{j=1}^{2}\prod_{k=1}^{d_j} P\left(\text{observed failure at time } t_{jk} \,\middle|\, R(t_{jk})\right) \\ &= \prod_{j=1}^{2}\prod_{k=1}^{d_j} \frac{\exp(b'z_{(jk)}(t_{jk}))}{\sum_{l\in R(t_{jk})}\exp(b'z_l(t_{jk}))}, \end{aligned}$$

where the risk set $R(t_{jk}) = \left\{js : \widetilde{T}_{js} \ge T_{jk}\right\}$ is the collection of subjects still on study just prior to t_{jk} in the j-th stage. Furthermore, the partial likelihood equation is

$$U(b) = \sum_{j=1}^{2}\sum_{k=1}^{d_j}\left[z_{(jk)}(t_{jk}) - e(b, t_{jk})\right] \tag{12.40}$$

where $e(b, t_{jk}) = \frac{\sum_{l\in R(t_{jk})}\exp(b'z_l(t_{jk}))z_l(t_{jk})}{\sum_{l\in R(t_{jk})}\exp(b'z_l(t_{jk}))}$. Based on (12.40), the corresponding observed information matrix is

$$I(b) = \sum_{j=1}^{2} \sum_{k=1}^{d_j} \left[\frac{\sum_{l \in R(t_{jk})} \exp(b' z_l(t_{jk})) z_l(t_{jk}) z_l'(t_{jk})}{\sum_{l \in R(t_{jk})} \exp(b' z_l(t_{jk}))} - e(b, t_{jk}) e'(b, t_{jk}) \right] \tag{12.41}$$

Test for Equality—Based on the formulation of the Cox model, the testing of equality can be conducted through the comparison of the coefficient vector b. Thus, consider the following hypotheses

$$H_0 : b = b_0 \quad \text{versus} \quad H_1 : b \neq b_0. \tag{12.42}$$

To test the above hypotheses, the score statistic $T_s = U'(b_0) I^{-1}(b_0) U(b_0)$ is considered. Under the null hypothesis H_0, T_s asymptotically follows a chi-squared distribution with p degrees of freedom where $p = \dim(b)$ (Cox and Hinklet 1974). Thus, H_0 is rejcted at an approximate α level of significance if $T_s > \chi_p^2(\alpha)$, where $\chi_p^2(\alpha)$ is the α-upper quantile of a chi-squared random variable with p degrees of freedom.

Consider the special case that the treatment indicator is the only covariate considered in the study. Let $z_{jk} = 1$ for the test treatment and $z_{jk} = 0$ for the control treatment. Then, the baseline $h(t|0)$ is the hazard in the control treatment and b is the log relative risk which measures the relative treatment effect. In particular, $b > 0$ (< 0) implies the test treatment increases (decreases) the risk of failure and $b = 0$ means no difference in risk between the two treatments. Define $P_{jk} = n_{Tjk} e^b / (n_{Tjk} e^b + n_{Rjk})$, where n_{Tjk} and n_{Rjk} denotes the number of subjects at risk, i.e., those who have not failed or censored just prior to the kth observed failure in the jth stage in the test and the control treatment, respectively. Consequently, the score function in (12.42) and the observed Fisher information matrix can be simplified to

$$U(b) = \sum_{j=1}^{2} \sum_{k=1}^{d_j} \left[z_{(jk)} - P_{jk} \right] \quad \text{and} \quad I(b) = \sum_{j=1}^{2} \sum_{k=1}^{d_j} P_{jk}(1 - P_{jk}),$$

respectively, where $z_{(jk)}$ is the treatment indicator for the kth observed failure in the jth stage. For the testing of the hypotheses of the equality of the two treatments defined in (12.42), the corresponding score test statistic is

$$T_s = \frac{U^2(0)}{I(0)} = \frac{\left[\sum_{j=1}^{2} \sum_{k=1}^{d_j} \left(z_{(jk)} - \frac{n_{Tjk}}{n_{Rjk} + n_{Tjk}} \right) \right]^2}{\sum_{j=1}^{2} \sum_{k=1}^{d_j} \frac{n_{Rjk} n_{Tjk}}{(n_{Rjk} + n_{Tjk})^2}}. \tag{12.43}$$

Under the null hypothesis in (12.43), T_s is asymptotically distributed as a chi-squared distribution with 1 degree of freedom. Equivalently, consider the statistic

$$T_z(b) = \frac{U(b)}{I^{1/2}(b)} = \frac{\sum_{j=1}^{2}\sum_{k=1}^{d_j}(z_{(jk)} - P_{jk})}{\sqrt{\sum_{j=1}^{2}\sum_{k=1}^{d_j} P_{jk}(1-P_{jk})}}. \tag{12.44}$$

$T_z(0)$ is asymptotic standard normal distributed. Therefore, the null hypothesis in (12.44) is rejected at an approximate α level of significance if $T_z(0) > z_{\alpha/2}$, where z_u is the upper u-quantile of the standard normal distribution.

Test for Superiority/Non-inferiority—Note that the log relative risk $b > 0$ implies worse treatment effect (inferiority) of the test treatment and $b < 0$ indicates better treatment effect (superiority) of the test treatment. In order to demonstrate superiority/non-inferiority, the following hypotheses are considered

$$H_0 : b \geq \delta \quad \text{versus} \quad H_1 : b < \delta, \tag{12.45}$$

where δ is a given superiority or non-inferiority margin. For $\delta < 0\,(> 0)$, the rejection of the null hypothesis implies superiority (non-inferiority) of the test treatment against the control. If $\delta - b$ is of order $O(n_1^{-1/2})$, then following similar arguments in Schoenfeld (1981), $T_z(b)$ is asymptotically normally distributed with unit variance and mean $\mu(b)$ given by

$$n_1^{1/2}(b-\delta)\left[\int_0^{cL} \pi_1(t,\delta)(1-\pi_1(t,\delta))V_1(t)dt + \rho\int_0^{L} \pi_2(t,\delta)(1-\pi_2(t,\delta))V_2(t)dt\right]^{1/2} \tag{12.46}$$

Consequently when $b = \delta$, the test statistic $T_z(b)$ approximately follows a standard normal distribution for sufficiently large sample size. Thus, the null hypothesis H_0 is rejected at an approximate level α of significance if $T_z(\delta) < -z_\alpha$.

Test for Equivalence—If the question of interest is to assess whether the performance of the test treatment is better than the (active) control, especially when prior knowledge of the test treatment is not available, it is more appropriate to consider the following hypotheses for the testing of therapeutic equivalence:

$$H_0 : |b| > \delta \quad \text{versus} \quad H_1 : |b| < \delta. \tag{12.47}$$

Since $|b| > \delta$ is equivalent to $b > \delta$ or $b < -\delta$. The above hypotheses can be tested by two one-sided test procedures. In particular, the null hypothesis is rejected at an approximate α level of significance if $T_z(\delta) < -z_\alpha$ and $T_z(-\delta) > z_\alpha$.

12.5 Analysis DS and DD Two-Stage Adaptive Designs

For a Category III DS two-stage adaptive design, the study objectives at different stages are different (e.g., dose selection versus efficacy confirmation) but the study endpoints are same at different stages. For a Category IV design, both study objectives

and endpoints at different stages are different (e.g., dose selection versus efficacy confirmation with surrogate endpoint versus clinical study endpoint).

As indicated earlier, how to control the overall type I error rate at a pre-specified level is one of the major regulatory concerns when adaptive design methods are employed in confirmatory clinical trials. Another concern is how to perform power analysis for sample size calculation/allocation for achieving individual study objectives originally set by the two separate studies (different stages). In addition, how to combine data collected from both stages for a combined and valid final analysis. Under a Category III or IV phase II/III seamless adaptive design, in addition, the investigator plans to have an interim analysis at each stage. Thus, if we consider the initiation of the study, first interim analysis, end of Stage 1 analysis, second interim analysis, and final analysis as critical milestones, the two-stage adaptive design becomes a 4-stage transitional seamless trial design. In what follows, we will focus on analysis of a four-stage transitional seamless design without (non-adaptive version) and with (adaptive version) adaptations, respectively.

12.5.1 Non-adaptive Version

For a given clinical trial comparing k treatments groups, $E_1, \ldots, E_k$ with a control group C, suppose a surrogate (biomarker) endpoint and a well-established clinical endpoint are available for assessment of the treatment effect. Denoted by θ_i and $\psi_i,\ i = 1, \ldots, k$ the treatment effect comparing E_i with C assessed by the surrogate (biomarker) endpoint and the clinical endpoint, respectively. Under the surrogate and clinical endpoints, the treatment effect can be tested by the following hypotheses:

$$H_{0,2} : \psi_1 = \cdots = \psi_k, \tag{12.48}$$

which is for the clinical endpoint, while the hypothesis

$$H_{0,1} : \theta_1 = \cdots \theta_k, \tag{12.49}$$

is for the surrogate (biomarker) endpoint. Cheng and Chow (2016) assumed that ψ_i is a monotone increasing function of the corresponding θ_i and proposed to test the hypotheses (12.48) and (12.49) at 3 stages (i.e., Stage 1, Stage 2a, Stage 2b, and Stage 3) based on accrued data at 4 interim analyses. Their proposed tests are briefly described below. For simplicity, he variances of the surrogate (biomarker) endpoint and the clinical outcome are denoted by σ^2 and τ^2, which are assumed known.

Stage 1—At this stage, $(k+1)n_1$ subjects are randomly assigned to receive either one of the k treatments or the control at a 1:1 ratio. In this case, we have n_1 subjects in each group. At the first interim analysis, the most effective treatment will be selected based on the surrogate (biomarker) endpoint and proceed to subsequent stages. For pairwise comparison, consider test statistics $\hat{\theta}_{i,1},\ i = 1, \ldots, k$ and $S = argmax_{1 \le j \le k} \hat{\theta}_{i,1}$. Thus, if $\hat{\theta}_{S,1} \le c_1$ for some pre-specified critical value c_1, then the

trial is stopped and we are in favor of $H_{0,1}$. On the other hand, if $\hat{\theta}_{S,1} > c_{1,1}$, then we conclude that the treatment E_S is considered the most promising treatment and proceed to subsequent stages. Subjects who receive either the promising treatment or the control will be followed for the clinical endpoint. Treatment assessment for all other subjects will be terminated but will undergo necessary safety monitoring.

Stage 2a—At Stage 2a, $2n_2$ additional subjects will be equally randomized to receive either the treatment E_S or the control C. The second interim analysis is scheduled when the short term surrogate measures from these $2n_2$ Stage 2 subjects and the primary endpoint measures from those $2n_1$ Stage 1 subjects who receive either the treatment E_S or the control C become available. Let $T_{1,1} = \hat{\theta}_{S,1}$ and $T_{1,2} = \hat{\psi}_{S,1}$ be the pair-wise test statistics from Stage 1 based on the surrogate endpoint and the primary endpoint, respectively, and $\hat{\theta}_{S,2}$ be the statistic from Stage 2 based on the surrogate. If

$$T_{2,1} = \sqrt{\frac{n_1}{n_1+n_2}}\hat{\theta}_{S,1} + \sqrt{\frac{n_2}{n_1+n_2}}\hat{\theta}_{S,2} \le c_{2,1},$$

then stop the trial and accept $H_{0,1}$. If $T_{2,1} > c_{2,1}$ and $T_{1,2} > c_{1,2}$, then stop the trial and reject both $H_{0,1}$ and $H_{0,2}$. Otherwise, if $T_{2,1} > c_{2,1}$ but $T_{1,2} \le c_{1,2}$, then we will move on to Stage 2b.

Stage 2b—At Stage 2b, no additional subjects will be recruited. The third interim analysis will be performed when the subjects in Stage 2a complete their primary endpoints. Let

$$T_{2,2} = \sqrt{\frac{n_1}{n_1+n_2}}\hat{\psi}_{S,1} + \sqrt{\frac{n_2}{n_1+n_2}}\hat{\psi}_{S,2},$$

where $\hat{\psi}_{S,2}$ is the pair-wise test statistic from Stage 2b. If $T_{2,2} > c_{2,2}$, then stop the trial and reject $H_{0,2}$. Otherwise, we move on to Stage 3.

Stage 3—At Stage 3, the final stage, $2n_3$ additional subjects will be recruited and followed till their primary endpoints. At the fourth interim analysis, define

$$T_3 = \sqrt{\frac{n_1}{n_1+n_2+n_3}}\hat{\psi}_{S,1} + \sqrt{\frac{n_2}{n_1+n_2+n_3}}\hat{\psi}_{S,2} + \sqrt{\frac{n_1}{n_1+n_2+n_3}}\hat{\psi}_{S,3},$$

where $\hat{\psi}_{S,3}$ is the pair-wise test statistic from Stage 3. If $T_3 > c_3$, then stop the trial and reject $H_{0,2}$; otherwise, accept $H_{0,2}$. The parameters in the above designs, $n_1,\ n_2,\ n_3, c_{1,1},\ c_{1,2},\ c_{2,1},\ c_{2,2}$, and c_3 are determined such that the procedure will have a controlled type I error rate of α and a target power of $1-\beta$.

In the above design, the surrogate data in the first stage are used to select the most promising treatment rather than assessing $H_{0,1}$. This means that upon completion of stage one a dose does not need to be significance in order to be used in subsequent stages. In practice, it is recommended that the selection criterion be based on precision analysis (desired precision or maximum error allowed) rather than power analysis (desired power). This property is attractive to the investigator since it does not suffer from any lack of power because of limited sample sizes.

As discussed above, under the 4-stage transitional seamless design, two sets of hypotheses, namely $H_{0,1}$ and $H_{0,2}$ are to be tested. Since the rejection of $H_{0,2}$ leads to the claim of efficacy, it is considered the hypothesis of primary interest. However, in the interest of controlling the overall type I error rate at a pre-specified level of significance, $H_{0,1}$ need to be tested following the principle of closed testing procedure to avoid any statistical penalties.

In summary, the two-stage phase II/III seamless adaptive design is attractive due to its efficiency, such as potentially reducing the lead time between studies (i.e., a phase II trial and a phase III study) and flexibility, such as making an early decision and taking appropriate actions (e.g. stop the trial early or delete/add dose groups.

12.5.2 Adaptive Version

The approach for trial design with non-adaptive version discussed in the previous section is basically a group sequential procedure with treatment selection at interim. There are no additional adaptations involved. With additional adaptations (adaptive version), Tsiatis and Metha (2003) and Jennison and Turnbull (2006) argue that adaptive designs typically suffer from loss of efficiency and hence are typically not recommended in regular practice. Proschan et al. (2006), however, also indicated that in some scenarios, particularly when there is not enough primary outcome information available, it is appealing to use an adaptive procedure as long as it is statistically valid and justified. The transitional feature of the multiple stage design enables us not only to verify whether the surrogate (biomarker) endpoint is predictive of the clinical outcome, but also to modify the design adaptively after the review of interim data. A possible modification is to adjust the treatment effect of the clinical outcome while validating the relationship between the surrogate (e.g. biomarker) endpoint and the clinical outcome. In practice, it is often assumed that there exists a local linear relationship between ψ and θ, which is a reasonable assumption if we focus only on the values at a neighborhood of the most promising treatment E_S. Thus, at the end of Stage 2a, we can re-estimate the treatment effect of the primary endpoint using

$$\hat{\delta}_S = \frac{\hat{\psi}_{S,1}}{\hat{\theta}_{S,1}} T_{2,1}.$$

Consequently, sample size can be re-assessed at Stage 3 based on a modified treatment effect of the primary endpoint $\delta = \max\{\delta_S, \delta_0\}$, where δ_0 is a minimally clinically relevant treatment effect. Suppose m is the re-estimated Stage 3 sample size based on δ. Then, there is no modification for the procedure if $m \leq n_3$. On the other hand, if $m > n_3$, then m (instead of n_3 as originally planned) subjects per arm will be recruited at Stage 3. The detailed justification of the above adaptation can be found in Cheng and Chow (2016).

12.5.3 A Case Study of Hepatitis C Virus Infection

A pharmaceutical company is interested in conducting a clinical trial for evaluation of safety, tolerability and efficacy of a test treatment for patients with hepatitis C virus (HCV) infection. For this purpose, a two-stage seamless adaptive design is considered. The proposed trial design is to combine two independent studies (one phase IIb study for treatment selection and one phase III study for efficacy confirmation) into a single study. Thus, the study consists of two stages: treatment selection (Stage 1) and efficacy confirmation (Stage 2). The study objective at the first stage is for treatment selection, while the study objective at Stage 2 is to establish the non-inferiority of the treatment selected from the first stage as compared to the standard of care (SOC). Thus, this is a typical Category IV design (a two-stage adaptive design with different study objectives at different stages).

For genotype 1 HCV patients, the treatment duration is usually 48 weeks of treatment followed by a 24 weeks follow-up. The well-established clinical endpoint is the sustained virologic response (SVR) at week 72. The SVR is defined as an undetectable HCV RNA level (<10 IU/mL) at week 72. Thus, it will take a long time to observe a response. The pharmaceutical company is interested in considering a biomarker or a surrogate endpoint such as a regular clinical endpoint with short duration to make early decision for treatment selection of four active treatments under study at end of Stage 1. As a result, the clinical endpoint of early virologic response (EVR) at week 12 is considered as a surrogate endpoint for treatment selection at Stage 1. At this point, the trial design has become a typical Category IV adaptive trial design (i.e., a two-stage adaptive design with different study endpoints and different study objectives at different stages). The resultant Category IV adaptive design is briefly outline below (Fig. 12.1):

Stage 1—At this stage, the design begins with five arms (4 active treatment arms and one control arm). Qualified subjects are randomly assigned to receive one of the five treatment arms at a 1:1:1:1:1 ratio. After all Stage 1 subjects have completed Week 12 of the study, an interim analysis will be performed based on EVR at week 12 for treatment selection. Treatment selection will be made under the assumption that the 12 week EVR is predictive of 72 week SVR. Under this assumption, the most promising treatment arm will be selected using precision analysis under some pre-specified selection criteria. In other words, the treatment arm with highest confidence level for achieving statistical significance (i.e., the observed difference as compared to the control is not by chance alone) will be selected. Stage 1 subjects who have not yet completed the study protocol will continue with their assigned therapies for the remainder of the planned 48 weeks, with final follow-up at Week 72. The selected treatment arm will then proceed to Stage 2.

Stage 2—At Stage 2, the selected treatment arm from Stage 1 will be test for non-inferiority against the control (SOC). A separate cohort of subjects will be randomized to receive either the selected treatment from Stage 1 or the control (SOC) at a 1:1 ratio. A second interim analysis will be performed when all Stage 2 subjects have completed Week 12 and 50% of the subjects (Stage 1 and Stage 2 combined)

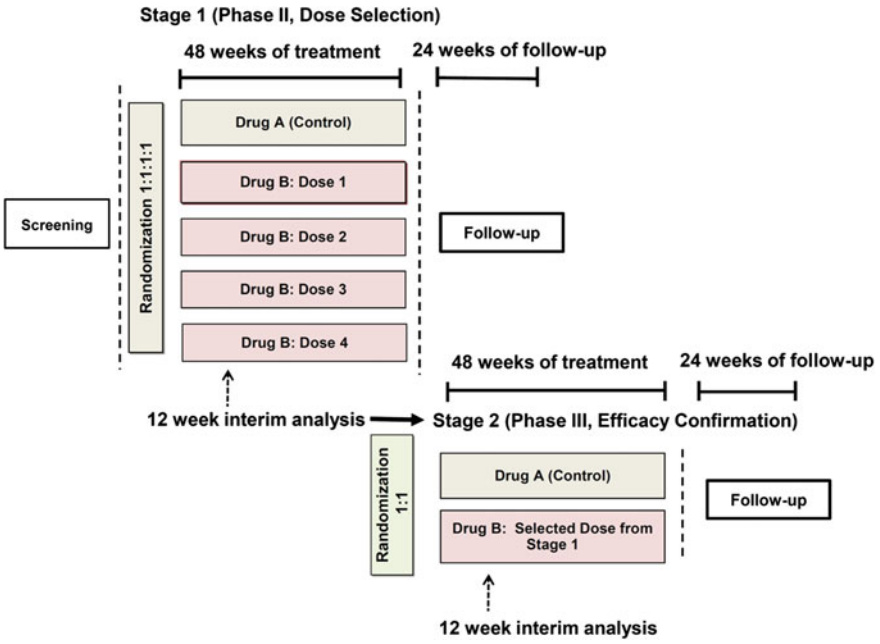


Fig. 12.1 A diagram of 4-stage transitional seamless trial design

have completed 48 weeks treatment and follow-up of 24 weeks. The purpose of this interim analysis is twofold. First, it is to validate the assumption that EVR at week 12 is predictive of SVR at week 72. Second, it is to perform sample size re-estimation to determine whether the trial will achieve study objective (establishing non-inferiority) with the desired power if the observed treatment preserves till the end of the study.

Statistical tests as described in the previous section will be used to test non-inferiority hypotheses at interim analyses and at end of stage analyses. For the two planned interim analyses, the incidence of EVR at week 12 as well as safety data, will be reviewed by an independent data safety monitoring board (DSMB). The commonly used O'Brien-Fleming type of conservative boundaries will be applied for controlling the overall Type I error rate at 5% (O'Brien and Fleming 1979). Adaptations such as stopping the trial early, discontinuing selected treatment arms, and re-estimating the sample size based on the pre-specified criteria may be applied as recommended by the DSMB. Stopping rules for the study will be designated by the DSMB, based on their ongoing analyses of the data and as per their charter.

12.6 Concluding Remarks

Chow and Chang (2011) pointed out that the standard statistical methods for a group sequential trial (with one planned interim analysis) is often applied for planning and data analysis of a two-stage adaptive design regardless whether the study objectives and/or the study endpoints are the same at different stages. As discussed earlier, two-stage seamless adaptive designs can be classified into four categories depending upon the study objectives and endpoints used at different stages. The direct application of standard statistical methods leads to the concern that the obtained p-value and confidence interval for assessment of the treatment effect may not be correct or reliable. Most importantly, sample size required for achieving a desired power obtained under a standard group sequential trial design may not be sufficient for achieving the study objectives under the two-stage seamless adaptive trial design, especially when the study objectives and/or study endpoints at different stages are different. Detailed information regarding sample size requirement for two-stage adaptive designs can be found in Chow et al. (2007).

As indicated in the 2010 FDA draft guidance on adaptive clinical trial design, adaptive designs were classified as either well understood designs or less well understood designs depending upon the availability of well-established statistical methods of specific designs (2010). In practice, most of the adaptive designs (including the two-stage seamless adaptive designs discussed in this article) are considered less well understood designs. Thus, the major challenge is not only the development of valid statistical methods for those less well understood designs, but also the development of a set of criteria for choosing an appropriate design among these less well understood designs for valid and reliable assessment of test treatment under investigation.

References

Bauer, P., & Kohne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics, 50,* 1029–1041.

Bauer, P., & Rohmel, J. (1995). An adaptive method for establishing a dose-response relationship. *Statistics in Medicine, 14,* 1595–1607.

Chang, M. (2007). Adaptive design method based on sum of p-values. *Statistics in Medicine, 26,* 2772–2784.

Cheng, B., & Chow, S. C. (2016). Statistical inference for a multiple-stage transitional seamless trials designs with different study objectives and endpoints.

Chow, S. C. (2011). *Controversial issues in clinical trials*. New York: Chapman and Hall/CRC, Taylor & Francis.

Chow, S. C., & Chang, M. (2011). *Adaptive design methods in clinical trials* (2nd ed.). New York: Chapman and Hall/CRC, Taylor and Francis.

Chow, S. C., Chang, M., & Pong, A. (2005). Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics, 15,* 575–591.

Chow, S. C., Lu, Q. S., & Tse, S. K. (2007). Statistical analysis for two-stage seamless design with different study endpoints. *Journal of Biopharmaceutical Statistics, 17,* 1163–1176.

Chow, S. C., Shao, J., Wang, H., & Lokhnygina, Y. (2017). *Sample size calculations in clinical research*. New York: Chapman and Hall/CRC Press, Taylor & Francis.
Chow, S. C., & Tu, Y. H. (2008). On two-stage seamless adaptive design in clinical trials. *Journal of the Formosan Medical Association, 107,* S52–S60.
Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of Royal Statistical Society B, 74,* 187–220.
Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman and Hall.
Cui, L., Hung, H. M. J., & Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics, 55,* 853–857.
FDA. (2010). Draft guidance for industry—adaptive design clinical trials for drugs and biologics. http://www.fda.gov/downloads/Drugs/…/Guidances/ucm201790.pdf.
Graybill, F. A., & Deal, R. B. (1959). Combining unbiased estimators. *Biometrics, 15,* 543–550.
Hommel, G., Lindig, V., & Faldum, A. (2005). Two-stage adaptive designs with correlated test statistics. *Journal of Biopharmaceutical Statistics, 15,* 613–623.
Jennison, C., & Turnbull, B. W. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika, 93,* 1–21.
Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. London/Boca Raton, FL: Chapman & Hall/CRC.
Khatri, C. G., & Shah, K. R. (1974). Estimation of location of parameters from two linear models under normality. *Communications in Statistics-Theory and Methods, 3,* 647–663.
Lehmacher, W., & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics, 55,* 1286–1290.
Li, G., Shih, W. C. J., & Wang, Y. N. (2005). Two-stage adaptive design for clinical trials with survival data. *Journal of Biopharmaceutical Statistics, 15,* 707–718.
Liu, Q., & Chi, G. Y. H. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics, 57,* 72–177.
Lu, Q. S., Tse, S. K., Chow, S. C., & Lin, M. (2012). Analysis of time-to-event data with non-uniform patient entry and loss to follow-up under a two-stage seamless adaptive design with Weibull distribution. *Journal of Biopharmaceutical Statistics, 22,* 773–784.
Meier, P. (1953). Variance of a weighted mean. *Biometrics, 9,* 59–73.
Muller, H. H., & Schafer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics, 57,* 886–891.
Obrien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical-trials. *Biometrics, 35,* 549–556.
Posch, M., & Bauer, P. (2000). Interim analysis and sample size reassessment. *Biometrics, 56,* 1170–1176.
Proschan, M. A., & Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics, 51,* 1315–1324.
Proschan, M. A., Lan, G. K. K., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. LLC, New York, NY: Springer Science + Business Media.
Rosenberger, W. F., & Lachin, J. M. (2003). *Randomization in clinical trials*. New York: Wiley.
Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika, 68,* 316–319.
Todd, S. (2003). An adaptive approach to implementing bivariate group sequential clinical trial designs. *Journal of Biopharmaceutical Statistics, 13,* 605–619.
Tsiatis, A. A., & Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika, 90,* 367–378.

Index

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2