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 3 Pharmaceutical Applications

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 3 Pharmaceutical Applications

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 and Gillings School of Global Public Health
University of North Carolina
Chapel Hill, NC, USA

ISSN 2199-0980 ISSN 2199-0999 (electronic)
ICSA Book Series in Statistics
ISBN 978-981-10-7819-4 ISBN 978-981-10-7820-0 (eBook)
https://doi.org/10.1007/978-981-10-7820-0

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 forty-six 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 19 chapters. The three volumes include the works of 70 authors or co-authors.

History of BASS: BASS was founded in 1994, by Dr. Karl E. Peace. He 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 Hsu College of Public Health, Georgia Southern University.

Originally, there were three objectives of BASS. Since the first editor founded the Journal of Biopharmaceutical Statistics (JBS) three 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, Georgia Southern University (GSU).

After 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, San Diego, CA; BASS IV–VII were held at the Hilton Oceanfront Resort Hotel, Palmetto Dunes, 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, 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 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's 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 FDA.

Volume III of the BASS Book Series, Entitled Pharmaceutical Applications, consists of 19 chapters. Chapter 1 presents targeted learning methods for determining optimal individualized treatment rules under cost constraints. Chapter 2 provides an overview of omics biomarker discovery and design considerations for biomarker-informed clinical trials. Chapter 3 presents methods for adaptive biomarker subpopulation and tumor type selection in Phase III clinical trials in oncology. Chapter 4 discusses high-dimensional data analysis methods in genomics. Chapter 5 provides an example of the importance of defining the primary endpoint drawing on properties of synergy or additivity. Chapter 6 presents the important method for the recycling of significance levels in testing multiple hypotheses in confirmatory clinical trials.

Chapter 7 then discusses the statistical testing of single and multiple endpoint hypotheses in **group sequential clinical trials**. Chapter 8 presents recently developed expanded statistical decision rules for interpretations of results of rodent carcinogenicity studies of pharmaceutical compounds. Chapter 9 provides a statistical analysis of a randomized, controlled clinical trial of Alendronate treatment

in Gaucher's (rare) disease that was prematurely halted. Chapter 10 discusses methods for mediation modeling in randomized trials in which the outcome variables are non-normal. Chapter 11 presents statistical considerations in using images obtained in clinical trials. Chapter 12 provides statistical analysis of several interesting applications arising from the statistical support of a variety of clinical trials.

In Chap. 13, important statistical consideration for uncovering fraud, misconduct, and other data quality issues in clinical trials is presented. Chapter 14 provides a thorough presentation of the essentials in the design and analysis of bio-similar studies. Chapter 15 presents a common language for causal estimands. Chapter 16 provides a thorough presentation of the development of prognostic biomarker signatures for survival using high-dimensional data. Chapter 17 develops methods for the validation, multivariate modeling, and the construction of heat map prediction matrices for overall survival in the context of missingness. Chapter 18 presents interesting biopharmaceutical applications using tepee plots. Finally, Chap. 19 presents methods for longitudinal and cross-sectional visualization with further applications in the context of heat maps.

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

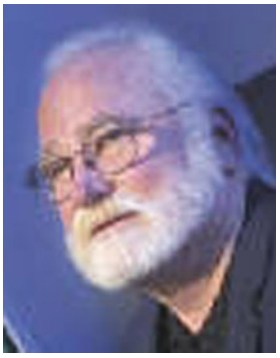
Chapel Hill, USA/Pretoria, South Africa
Ding-Geng Chen, Ph.D.
Professor, University of North Carolina
Extraordinary Professor, University of Pretoria

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), Georgia Southern University, Statesboro, GA. He is Fellow of the American Statistical Association (ASA), Founding Editor of the Journal of Biopharmaceutical Statistics, Founding Director of the Center for Biostatistics in JPHCOPH, Founder of BASS, and Endower of JPHCOPH. He is the recipient of numerous awards and citations from 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 15 books and over 100 publications.

Prof. Ding-Geng Chen is 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 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 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 Vice President and 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 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 Drug Information Association (DIA); 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 Associate Editor of American Statistical Association (ASA) journal *Statistics in Biopharmaceutical Research* (SBR) and as Selection Committee Member of *Samuel S. Wilks Memorial Award* offered by ASA.

Contributors

Mohamed Alosh Division of Biometrics III, Office of Biostatistics, OTS, CDER, FDA, Silver Spring, MD, USA

Dhammika Amaratunga Princeton Data Analytics LLC, Bridgewater, NJ, USA

Sibabrata Banerjee Department of Biostatistics, Celgene Corporation, Summit, NJ, USA

Robert A. Beckman Departments of Oncology and of Biostatistics, Bioinformatics, and Biomathematics, Lombardi Comprehensive Cancer Center and Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington, DC, USA

Javier Cabrera Department of Statistics, Rutgers University, Piscataway, NJ, USA

Cong Chen Biostatistics and Research Decision Sciences, Merck & Co, Inc., Kenilworth, NJ, USA

Jing Cheng University of California, San Francisco, USA

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

Albert Elion-Mboussa Department of Biostatistics, Celgene Corporation, Summit, NJ, USA

Stuart A. Gansky University of California, San Francisco, USA

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

Mia He Department of Statistical Programming, Celgene Corporation, Summit, NJ, USA

Bo Huang Global Product Development, Pfizer Inc., Groton, CT, USA

Mohammad Huque Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, USA

Vatsala Karwe Department of Biostatistics, Celgene Corporation, Summit, NJ, USA

Stanley Kotey Department of Biostatistics, Celgene Corporation, Summit, NJ, USA

Wen Li Biostatistics and Research Decision Sciences, Merck & Co, Inc., Kenilworth, NJ, USA

Xiaoyun (Nicole) Li Biostatistics and Research Decision Sciences, Merck & Co, Inc., Kenilworth, NJ, USA

Karl K. Lin Division of Biometrics 6, Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA

Ying Lu Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA

Sandeep Menon Worldwide Research and Development, Pfizer Inc., Cambridge, MA, USA

Sirisha Mushti Division of Biometrics V, Office of Biostatistics, OTS, CDER, FDA, Silver Spring, MD, USA

Uche Eseoghene Okoro Jiann-Ping Hsu School of Public Health, Georgia Southern University, Statesboro, GA, USA

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

Mohammad A. Rahman Division of Biometrics 6, Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA

Bruce E. Rodda Strategic Statistical Consulting LLC, University of Texas School of Public Health, Austin, TX, USA

Richard Simon R Simon Consulting, Rockville, MD, USA

Fuyu Song Center for Food and Drug Inspection, China Food and Drug Administration, Beijing, China

Rick Soong Department of Statistical Programming, Celgene Corporation, Summit, NJ, USA

Shankar S. Srinivasan Department of Biostatistics, Celgene Corporation, Summit, NJ, USA

Shumei S. Sun Virginia Commonwealth University, Richmond, VA, USA

Ye Tan Early Clinical Development, Pfizer Inc., Cambridge, MA, USA

Boriska Toth UC-Berkeley, Berkeley, USA

Kao-Tai Tsai Jiann-Ping Hsu School of Public Health, Georgia Southern University, Statesboro, GA, USA

Mark van der Laan UC-Berkeley, Berkeley, USA

Jing Wang Global Product Development, Pfizer Inc., Cambridge, MA, USA

Li Hua Yue Department of Biostatistics, Celgene Corporation, Summit, NJ, USA

Weidong Zhang Global Product Development, Pfizer Inc., Cambridge, MA, USA

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

Chapter 1
Targeted Learning of Optimal Individualized Treatment Rules Under Cost Constraints

Boriska Toth and Mark van der Laan

1.1 Introduction

We consider a general resource-allocation problem, namely, to maximize a mean outcome given a cost constraint, through the choice of a treatment rule that is a function of an arbitrary fixed subset of an individual's covariates. In pharmaceutical applications, we typically think of maximizing a clinical outcome given a monetary cost constraint, through the allocation of medication to patients, although our model is much more general. We focus on the setting where unmeasured confounding is a possibility, but a valid instrumental variable is available. Thus, our setup allows for consistent estimation of the optimal treatment rule and causal effects in a range of non-randomized studies, including post-market and other observational studies, as well as studies involving imperfect randomization due to non-adherence. The goal is both to: (1) find an optimal intervention $d(V)$ for maximizing the mean counterfactual outcome, where V is an arbitrary fixed subset of baseline covariates W, and (2) estimate the mean counterfactual outcome under this rule $d(V)$. We make no restrictions on the type of data; however, the case of a continuous or categorical instrument or treatment variable is discussed in Toth (2016). To our knowledge, this work is the first to estimate the effect of an optimal individualized treatment regime, under a non-unit cost constraint, in the instrumental variables setting.

Utilizing instrumental variables. A classic solution for obtaining a consistent estimate of a causal effect under unmeasured confounding is to use an instrumental variable, assuming one exists. Informally, an instrumental variable, or instrument, is a variable Z that affects the outcome Y only through its effect on the treatment A, and the residual (error) term of the instrument is uncorrelated with the residual term of the

B. Toth (✉) · M. van der Laan
UC-Berkeley, Berkeley, USA
e-mail: bori@stat.berkeley.edu

M. van der Laan
e-mail: laan@berkeley.edu

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_1

outcome (Imbens and Angrist 1994; Angrist et al. 1996; Angrist and Krueger 1991). Thus, the instrument produces exogenous variation in the treatment. Instrumental variables have been used widely in biostatistics and pharmaceutics. (See Brookhart et al. 2010 for a large collection of references.) In these settings, the instrumental variable is usually some attribute that is related to the health care a patient receives, but is not at the level of individual patients. For example, Brookhart and Schneeweiss (2007) exploit variation in physician preference for prescribing NSAID medications to infer the effect of these medications on gastrointestinal bleeding.

In this work, we solve two versions of the optimal individualized treatment problem: (1) when the intervention is on the treatment variable A (Sect. 1.7), and (2) when the intervention is actually on the instrument Z (Sect. 1.6). For example, consider a study in which HIV-positive patients were encouraged to undergo antiretroviral therapy (ART) with a randomized (or quasi-randomized) encouragement design, but a number of factors caused non-adherence among some patients (Chesney 2006). The methods in this chapter allow one to infer what would be the optimal assignment of patients to ART treatment, based on patient characteristics, to achieve a desirable outcome (i.e., suppressed viral load, 5-year survival), given a limited budget. One parameter of interest is the mean outcome under optimal assignment of individuals to actually receive ART. This is the problem of finding an optimal treatment regime. However, in this setting of non-adherence, it might not be possible to intervene directly on the treatment variable. Thus, another parameter of interest is the mean outcome under the optimal intervention on the instrumental variable. We call this the problem of finding an optimal *intent-to-treat* regime, so named because the instrument is often a randomized assignment to treatment or encouragement mechanism. Under our randomization assumption on instrument Z, the optimal intent-to-treat problem is the same as an optimal treatment problem without unmeasured confounding, as Z can be seen as a treatment variable that is unconfounded with Y.

Causal effects given arbitrary subgroups of the population.
A key feature of our work is that the optimal intervention $d(V)$ is a function of a fixed arbitrary subset V of all baseline covariates W. There is currently great interest and computational feasibility in designing individualized treatment regimes based on a patient's characteristics and biomarkers. The paradigm of precision medicine calls for incorporating high-dimensional spaces of genetic, environmental, and lifestyle variables into treatment decisions (Editors: National Research Council Committee 2011). Incorporating many covariates for estimating relevant components of the data-generating distribution can be helpful in: (1) improving the precision of the statistical model and (2) ensuring that the instrument induces exogenous variation given the covariates. However, a physician typically has a smaller set of patient variables that are available and that he/she considers reliable predictors. Thus, being able to calculate an optimal treatment (or intent-to-treat) regime as a function of an arbitrary subset of baseline covariates is of great use.

The targeted minimum loss-based framework.
Our estimators use targeted minimum loss-based estimation (TMLE), which is a methodology for semiparametric estimation that has very favorable theoretical

properties and can be superior to other estimators in practice (van der Laan and Rubin 2006; van der Laan and Rose 2011). TMLE guarantees asymptotic efficiency when certain components of the data-generating distribution are consistently estimated. Thus, under certain conditions, the TMLE estimator is optimal in having the asymptotically lowest variance for a consistent estimator in a general semiparametric model, thereby achieving the semiparametric Cramer–Rao lower bound (Newey 1990). The TMLE method also has a robustness guarantee: It produces consistent estimates even when the functional form is not known for all relevant components of the parameter of interest (see Sects. 1.6.3.4 and 1.7.3). Another beneficial property is asymptotic linearity. This ensures that TMLE-based estimates are close to normally distributed for moderate sample sizes, which makes for accurate coverage of confidence intervals. Finally, TMLE has the advantage over other semiparametric efficient estimators that it is a substitution estimator, meaning that the final estimate is made by evaluating the parameter of interest on the estimates of its relevant components. This property has been linked to good performance in sparse data in Gruber and van der Laan (2010).

The TMLE methodology uses the following procedure for constructing an estimator:

1. Let P_0 denote the true data-generating distribution. One first notes that the parameter of interest $\Psi(P_0)$ depends on P_0 only through certain relevant components Q_0 of the full distribution P_0; in other words, $\Psi(P_0) = \Psi(Q_0)$.[1] TMLE *targets* these relevant components by only estimating these Q_0 and certain nuisance parameters g_0[2] that are needed for updating the relevant components. An initial estimate (Q_n^0, g_n) is formed of the relevant components and nuisance parameters. This is typically done using the Super Learner approach described in van der Laan et al. (2007), in which the best combination of learning algorithms is chosen from a library using cross-validation.
2. Then, the relevant components Q_n^0 are fluctuated, possibly in an iterative process, in an optimal direction for removing bias efficiently. To do so, one defines a fluctuation function $\varepsilon \to Q(\varepsilon|g_n)$ and a loss function $L(\dots)$, where we fluctuate Q_n^0 to $Q_n^0(\varepsilon|g_n)$ by solving for fluctuation $\varepsilon = \text{argmin}_\varepsilon \frac{1}{n}\sum_{i=1}^n L(Q_n^0(\varepsilon|g_n), g_n)(O_i)$. For example, the loss function might be the mean squared error or the negative log likelihood function.
3. Finally, one evaluates the statistical target parameter on the updated relevant components Q_n^* and arrives at estimate $\psi_n^* = \Psi(Q_n^*)$.

The key requirement is to choose the fluctuation and loss functions so that, upon convergence of the components to their final estimate Q_n^* and g_n^*, the efficient influence curve equation is solved:

$$P_n\, D^*(Q_n^*, g_n^*) = 0$$

[1] We are abusing notation here for the sake of convenience by using $\Psi(\cdot)$ to denote the mapping both from the full distribution to $\mathbb{R}^d$ and from the relevant components to $\mathbb{R}^d$.

[2] The nuisance parameters are those components g_0 of the efficient influence curve $D^*(Q_0, g_0)$ that $\Psi(Q_0)$ does not depend on.

Above, P_n denotes the empirical distribution $(O_1, \ldots, O_n)$, and we use the shorthand notation $P_n f = \frac{1}{n}\sum_{i=1}^{n} f(O_i) \cdot D^*$ denotes the efficient influence curve.

1.2 Prior Work

Luedtke and van der Laan (2016a) is a recent work that gives a TMLE estimator for the mean outcome under optimal treatment given a cost constraint. That problem is very similar to the one we solve in Sect. 1.6, with the main difference being that we allow a more general non-unit cost constraint which results in a different closed-form solution to the optimal rule. Luedtke and van der Laan (2016b) tackles the issue of possible non-unique solutions and resulting violations of pathwise differentiability. The conditions we require in assumptions (A2)–(A4) are adopted from these works.

A large body of work focuses on the case of optimal treatment regimes in the unconstrained case, such as Robins (2004). More recently, various approaches tackle the constrained ODT problem: Zhang et al. (2012) describe a solution that assumes the optimal treatment regime is indexed by a finite-dimensional parameter, while Chakraborty et al. (2013) describe a bootstrapping method for learning ODT regimes with confidence intervals that shrink at a slower than root-n rate. Chakraborty and Moodie (2013) give a review of recent work on the constrained case.

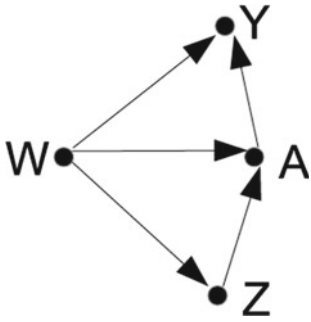
1.3 Model and Problem

We consider the problem of estimation and inference under an optimal intervention, in the context of an instrumental variable model. We take an iid sample of n data points $(W, Z, A, Y) \sim \mathscr{M}$, where $\mathscr{M}$ is a semiparametric model. Z is assumed to be a valid instrument for identifying the effect of treatment A on outcome Y, when one has to account for unmeasured confounding. In applications, instrument Z is often a randomized encouragement mechanism or randomized assignment to treatment which may or may not be followed. In other cases, Z is not perfectly randomized but nevertheless promotes or discourages individuals in receiving treatment. $V \subseteq W$ is an arbitrary fixed subset of the baseline covariates, and $F_V(W)$ gives the mapping $W \to V \cdot d(V)$ refers to a decision rule as a function of V, where $Z = d(V)$ is used to denote the optimal intervention on the instrument Z, in other words, the optimal assignment to treatment or the optimal intent-to-treat. $A = d(V)$ refers to the optimal treatment rule. We are interested in estimating the mean counterfactual outcome under an optimal rule $Z = d(V)$ or $A = d(V)$. Figure 1.1 shows a diagram.

There are no restrictions on the type of data. However, the case of categorical or continuous Z or A are both dealt with separately in Toth (2016).

Further, we let $c_A(A, W)$ be a cost function that gives the cost associated with assigning an individual with covariates W to a particular A value. We let $c_T(Z, W)$ be a cost function that gives the total cost associated with assigning an individual

Fig. 1.1 Causal diagram



with covariates W to a particular Z value. We can think of $c_T(Z, W)$ as the sum of $c_Z(Z, W)$, a cost incurred directly from setting Z, and $E_{A|W,Z}c_A(A, W)$, an average cost incurred from the actual treatment A.[3] We need to find optimal rule $Z = d(V)$ under cost constraint $E\ c_T(Z, W) \leq K$, for a fixed cost K, and optimal rule $Z = d(V)$ under constraint $E\ c_A(A, W) \leq K$.

Notation. Let $P_W \equiv \Pr(W)$ and $\rho(Z, W) \equiv \Pr(Z = 1|W)$. Also let $\Pi(Z, W) \equiv E(A \mid Z, W)$ be the conditional mean of A given Z, W, and $\mu(Z, W) \equiv E(Y \mid Z, W)$.

We also define $\mu_b(V) \triangleq E_{W|V}\big[\mu(Z = 1, W) - \mu(Z = 0, W)\big]$, which gives the mean difference in outcome between setting $Z = 1$ and $Z = 0$ given V. Similarly, $c_{b,Z}(V) \triangleq E_{W|V}\big[c_T(Z = 1, W) - c_T(Z = 0, W)\big]$, and $c_{b,A}(V) \triangleq E_{W|V}\big[c_A(A = 1, W) - c_A(A = 0, W)\big]$. We also use notation $m(V) \triangleq E_{W|V}m(W)$, where m is the causal effect function defined in the causal assumptions.

We further assume wlog that intent-to-treat $Z = 0$ has lower cost for all V: $E_{W|V}c_T(0, W) \leq E_{W|V}c_T(1, W)$.[4] Let $\underline{K_Z} \triangleq E_W c_T(0, W)$ be the total cost of not assigning any individuals to intent-to-treat, and $\overline{K_Z} \triangleq E_W c_T(1, W)$ be the total cost of assigning everyone, and we assume a non-trivial constraint $\underline{K_Z} < K < \overline{K_Z}$. Define $\underline{K_A} \triangleq E_W c_A(0, W)$, and $\overline{K_A}$ similarly.

Causal model.
Using the structural equation framework of (Pearl 2000), we assume that each variable is a function of other variables that affect it and a random term (also called error term). Let U denote the error terms. Thus, we have

$$W = f_W(U_W), Z = f_Z(W, U_Z), A = f_A(W, Z, U_A), Y = f_Y(W, Z, A, U_Y)$$

[3]It is not hard to extend this model to incorporate uncertainty in $E(A|W, Z)$ for calculating $c_T(Z, W)$, and thus estimating $c_T(Z, W)$ from the data, given fixed functions c_Z, c_A. There is a correction term that gets added to the efficient influence curve.

[4]We are only making this assumption for the sake of easing notation. We can forgo this assumption by introducing notation; i.e., $Z = l(V)$ is the lower cost intent-to-treat value for a stratum defined by covariates V.

where $U = (U_W, U_Z, U_A, U_Y) \sim P_{U,0}$ is an exogenous random variable, and f_W, f_Z, f_A, f_Y may be unspecified or partially specified (for instance, we might know that the instrument is randomized). U_Y is possibly confounded with U_A.

We use notation that a subscript of 0 denotes the true distribution, in expressions such as E_0, P_0.

Assumption (A1) ***Assumptions ensuring that Z is a valid instrument***:

1. ***Exclusion restriction***. Z only affects outcome Y through its effect on treatment A. Thus, $f_Y(W, Z, A, U_Y) = f_Y(W, A, U_Y)$.
2. ***Exogeneity of the instrument***. $E(U_Y|W, Z) = 0$ for any W, Z.
3. Z ***induces variation in*** A. $\text{Var}_0[E_0(A|Z, W)|W] > 0$ for all W.

Structural equation for outcome Y:

4. $Y = Am(W) + \theta(W) + U_Y$ for continuous Y, and
   $\Pr(Y = 1|W, A, \tilde{U}_Y) = Am(W) + \theta(W) + \tilde{U}_Y$ for binary Y,
   where $U_Y = (\tilde{U}_Y, U'_Y)$ for an exogenous r.v. U'_Y,[5] and m, θ are unspecified functions.

Assumptions 2 and 4 yield that, whether Y is binary or continuous,

$$E(Y|W, Z) = m_0(W)\Pi_0(W, Z) + \theta_0(W)$$

We use $Y(A = a)$ to denote the counterfactual from setting treatment to $A = a$. These assumptions guarantee that $E(Y(A = a))$ equals $E_W m(W)a + \theta(W)$ for identifiable functions m, θ.

It should be noted that we do not require the instrument to be randomized with respect to treatment ($U_Z \perp\!\!\!\perp U_A|\ W$ is not necessary).

It is simple to see from the above instrumental variable assumptions that Z is randomized with respect to Y, so we have:

Corollary 1 (*Randomization of Z.*) $U_Z \perp U_Y|W$.

This implies $E(Y(Z)|W) = E(Y|W, Z)$.

Statistical model. The above-stated causal model implies the statistical model $\mathcal{M}$ consisting of all distributions P of $O = (W, Z, A, Y)$ satisfying $E_P(Y|W, Z) = m_P(W) \cdot \Pi_P(W, Z) + \theta_P(W)$. Here, m_P and θ_P are unspecified functions and $\Pi_P(W, Z) = E_P(A|W, Z)$ such that $Var_P(\Pi_P(Z, W)|W) > 0$ for all W. Note that the regression equation $E_P(Y|W, Z) = m_P(W) \cdot \Pi_P(W, Z) + \theta_P(W)$ is always satisfied for some choice of $m(W)$, $\theta(W)$ when Z is binary. The distribution for the instrument $\rho(W)$ may or may not be known, and we generally think of all other components P_W, Π, m, θ as unspecified.

[5] The U'_Y term is an exogenous r.v. whose purpose is for sampling binary Y with mean $\tilde{f}_Y(W, Z, A, \tilde{U}_Y)$.

1.3.1 Parameter of Interest, with Optimal Intent-to-Treat

Causal parameter of interest.

$$\Psi_Z(P_0) \triangleq \text{Max}_d \; E_{P_0} Y(Z = d(V)) \text{ s.t. } E_{P_0}[c_T(Z = d(V), W)] \leq K$$

Statistical target parameter.

$$\Psi_{Z,0} = E_{P_0}\mu_0(Z = d_0(V), W) \tag{1.1}$$

where d_0 is the optimal intent-to-treat rule:
$d_0 = \text{argmax}_d \; E_{P_0}\mu_0(Z = d(V), W)$ s.t. $E_{P_0}[c_T(Z = d(V), W)] \leq K$
We also use the notation $\Psi_Z(P_0) = \Psi_Z(P_{W,0}, \mu_0)$.

1.3.2 Parameter of Interest, with Optimal Treatment

Causal parameter of interest.

$$\Psi_A(P_0) \triangleq \text{Max}_d \; E_0 Y(A = d(V)) \text{ s.t. } E_0[c_A(A = d(V), W)] \leq K \tag{1.2}$$

Identifiability. $m(W)$ is identified as $\big[(\mu(Z = 1, W) - \mu(Z = 0, W))/(\Pi(Z = 1, W) - \Pi(Z = 0, W))\big]$. $\theta(W)$ is identified as $\big[\mu(Z, W) - \Pi(Z, W) \cdot m(W)\big]$.

Statistical target parameter.

Lemma 1 *The causal parameter given in Eq. (1.2) is identified by the statistical target parameter:*

$$\Psi_{A,0} = E_{P_{W,0}}\big[m_0(W)d_0(V) + \theta_0(W)\big] \tag{1.3}$$

Note that optimal decision rule d_0 is a function of m_0, $P_{W,0}$. For $\Psi_{A,0}$ we also use the notation $\Psi_A(P_{W,0}, m_0, \theta_0)$, or alternately $\Psi_A(P_{W,0}, \Pi_0, \mu_0)$, using the above identifiability results.

This lemma follows from our causal assumptions:

$\Psi_A(P_0) = EY(A = d_0(V)) = E_W E_{U_Y|W} EY(A = d_0(V)|W, U_Y)$
The right hand side becomes $E_W E_{U_W|Y}(m(W)d_0(V) + \theta(W) + U_Y)$ for a continuous Y, and $E_W E_{U_W|Y}(m(W)d_0(V) + \theta(W) + \tilde{U}_Y)$ for a binary Y.

1.4 Closed-Form Solution for Optimal Rule d_0 in the Case of Binary Treatment

The problem of finding the optimal deterministic treatment rule $d(V)$ is NP-hard (Karp 1972). However, when allowing possible non-deterministic treatments, there is a simple closed-form solution for the optimal treatment or the optimal intent-to-treat. The optimal rule is to treat all strata with the highest marginal gain per marginal cost, so that the total cost of the policy equals the cost constraint.

This section introduces key quantities and notation used in the rest of the chapter. We present the solution in detail for the case of intervening on the instrument, when $Z = d_0(V)$. Recall that wlog we think of $Z = 0$ as the 'baseline' intent-to-treat (ITT) value having lower cost. We define a scoring function $T(V) = \frac{\mu_b(V)}{c_b(V)}$ for ordering subgroups (given by V) based on the effect of setting $Z = 1$ per unit cost. In the optimal intent-to-treat policy, all groups with the highest $T(V)$ values deterministically have Z set to 1, up to cost K and assuming $\mu_b \geq 0$. We write $T_P(V)$ to make explicit the dependence on P_W, $\mu(Z, W)$ from distribution P.

Define a function S_P: $[-\infty, +\infty] \to \mathbb{R}$ as

$$S_P(x) = E_V[I(T_P(V) \geq x)(c_b(V)]$$

In other words, $S_P(x)$ gives the expected (additional above baseline) cost of setting $Z = 1$ for all subgroups having $T_P(V) \geq x$. We use $S_0(\cdot)$ to denote S_{P_0} from here on.

Define cutoff η_P as

$$\eta_P = S_P^{-1}(K - \underline{K_{A,P}})$$

The assumptions below in Sect. 1.5 guarantee that $S_P^{-1}(K - \underline{K_{A,P}})$ exists and η_P is well defined. η is set so that there is a total cost K of treating with $Z = 1$ everyone having $T(V) \geq \eta$. Further let:

$$\tau_P = \max\{\eta_P, 0\}$$

Thus, τ gives the cutoff for the scoring function $T(V)$, so the optimal rule is

$$d_P(V) = 1 \text{ iff } T_P(V) \geq \tau_P$$

Lemma 2 *Assume (A2)–(A4). Then, the optimal decision rule d_0 for parameter $\Psi_{Z,0}$ as defined in Eq. 1.1 is the deterministic solution $d_0(V) = 1$ iff $T_0(V) \geq \tau_0$, with T_0, τ_0 as defined above.*

The proof is given in Toth (2016). That work also describes modifications to the optimal solution for d_0 when Z is continuous or categorical.

1.4.1 Closed-Form Solution for Optimal Treatment Rule $A = d_0(V)$

The solution given above goes through for the case of intervening on the treatment, with the two main modifications that: (1) replace intervention variable Z with A, and (2) replace $\mu_b(W)$ with $m(W)$. These latter quantities represent the effect on Y of applying the intervention versus the baseline treatment (at Z or A, respectively).

1.5 Assumptions for Pathwise Differentiability of $\Psi_{Z,0}$ and $\Psi_{A,0}$

We use notation $d_0 = d_{P_0}$, $\tau_0 = \tau_{P_0}$, etc. We state these assumptions for $\Psi_{Z,0}$. The exact same assumptions apply for $\Psi_{A,0}$, replacing Z with A in a few places.

These three assumptions are needed to ensure pathwise differentiability and prove the form of the canonical gradient (Theorem 1).

Assumptions (A2)–(A4).

(***A2***) Positivity assumption: $0 < \rho_0(W) < 1$.

(***A3***) There is a neighborhood of η_0 where $S_0(x)$ is Lipschitz continuous, and a neighborhood of $S_0(\eta_0) = K - \underline{K_Z}_0$ where $S_0^{-1}(y)$ is Lipschitz continuous.

(***A4***) $Pr_0(T_0(V) = \tau) = 0$ for all τ in a neighborhood of τ_0.

Note that (A3) implies that $S_0^{-1}(K - \underline{K_Z}_0)$ exists. Note also that (A3) actually implies $Pr_0(T_0(V) = \eta) = 0$ for η in a neighborhood of η_0, and thus, (A3) implies (A4) when $\eta_0 > 0$ and $\tau_0 = \eta_0$.

Need for (A4) (Guarantee of non-exceptional law).
If (A4) does not hold and there is positive probability of individuals being at the threshold for being treated or not under the optimal rule, then the solution $d(V)$ is not unique, and $\Psi_{Z,0}$ is no longer pathwise differentiable. It is easy to see that under (A4), the optimal $d(V)$ over the broader set of non-deterministic decision rules is a deterministic rule. Toth (2016) describes why (A4) is a reasonable assumption in practice when we have a constraint $\underline{K_Z} < K < \overline{K_Z}$ that allows for only a strict subset of the population to be treated.

1.6 TMLE for Optimal Intent-to-Treat Problem ($\Psi_{Z,0}$)

All proofs and derivations for what follows are given in Toth (2016).

1.6.1 Canonical Gradient for $\Psi_{Z,0}$

For $O = (W, Z, A, Y)$, and deterministic rule $d(V)$, define

$$D_1(d, P)(O) \triangleq \frac{I(Z = d(V))}{\rho_P(W)}(Y - \mu_P(Z, W)) \tag{1.4}$$

$$D_2(d, P)(O) \triangleq \mu_P(d(V), W) - E_P\mu_P(d(V), W) \tag{1.5}$$

$$D_3(d, \tau, P)(O) = -\tau(c_T(d(V), W) - K) \tag{1.6}$$

Define

$$D^*(d, \tau, P)(O) \triangleq D_1(d, P)(O) + D_2(d, P)(O) + D_3(d, \tau, P)(O)$$

Theorem 1 *Assume (A1)–(A4) above. Then Ψ_Z is pathwise differentiable at P_0 with canonical gradient $D_0 = D^*(d_0, \tau_0, P_0)$.*

1.6.2 TMLE

The relevant components for estimating $\Psi_Z = E_W\mu(Z = d(V), W)$ are $Q = (P_W, \mu(Z, W))$. Decision rule d is also part of Ψ, but it is a function of P_W, $\mu(Z, W)$. The nuisance parameter is $g = \rho(W)$. First convert Y to the unit interval via a linear transformation $Y \to \tilde{Y}$, so that $\tilde{Y} = 0$ corresponds to $Y_{\min}$ and $\tilde{Y} = 1$ to $Y_{\max}$. We assume $Y \in [0, 1]$ from here.

1. Use the empirical distribution $P_{W,n}$ to estimate P_W. Make initial estimates of $\mu_n(Z, W)$ and $g_n = \rho_n(W)$ using any strategy desired. Data-adaptive learning using Super Learner is recommended.
2. The empirical estimate $P_{W,n}$ gives an estimate of $Pr_{V,n}(V) = E_{W,n}I(F_V(W) = V)$, $\underline{K_{Z,n}} = E_{W,n}c_T(0, W)$, $\overline{K_{Z,n}} = E_{W,n}c_T(1, W)$, and $c_{b,Z,n}(V) = E_{W,n|V}(c_T(\overline{1, W}) - c_T(0, W))$.
3. Estimate $\mu_{b,0}$ as $\mu_{b,n}(V) = E_{W,n|V}(\mu_n(1, W) - \mu_n(0, W))$.
4. Estimate $T_0(V)$ as $T_n(V) = \frac{\mu_{b,n}(V)}{c_{b,Z,n}(V)}$.
5. Estimate $S_0(x)$ using $S_n(x) = E_{V,n}[I(T_n(V) \geq x)(c_{b,Z,n}(V)]$.
6. Estimate η_0 as η_n using $\eta_n = S_n^{-1}(K - \underline{K_{Z,n}})$ and $\tau_n = \max\{0, \eta_n\}$.

7. Estimate the decision rule as $d_n(V) = 1$ iff $T_n(V) \geq \tau_n$.
8. Now fluctuate the initial estimate of $\mu_n(Z, W)$ as follows: For $Z \in [0, 1]$, define covariate $H(Z, W) \triangleq \frac{I(d_n(V)=Z)}{g_n(W)}$. Run a logistic regression using:

   Outcome: $(Y_i : i = 1, \ldots, n)$
   Offset: $(\text{logit}\, \mu_n(Z_i, W_i), i = 1, \ldots, n)$
   Covariate: $(H(Z_i, W_i) : i = 1, \ldots, n)$

   Let ε_n represent the level of fluctuation, with $\varepsilon_n = \text{argmax}_\varepsilon \frac{1}{n} \sum_{i=1}^{n} [\mu_n(\varepsilon)(Z_i, W_i) \log Y_i + (1 - \mu_n(\varepsilon)(Z_i, W_i)) \log(1 - Y_i)]$ and $\mu_n(\varepsilon)(Z, W) = \text{logit}^{-1}(\text{logit}\, \mu_n(Z, W) + \varepsilon H(Z, W))$.
9. Set the final estimate of $\mu(Z, W)$ to $\mu_n^*(Z, W) = \mu_n(\varepsilon_n)(Z, W)$.
10. Finally, form final estimate of $\Psi_{Z,0} = \Psi_{Z,d_0}(P_0)$ using the plug-in estimator

$$\Psi_Z^* = \Psi_{Z,d_n}(P_n^*) = \frac{1}{n} \sum_{i=1}^{n} \mu_n^*(Z = d_n(V_i), W_i)$$

We have used the notation $\Psi_{Z,d}(P)$ referring to mean outcome under decision rule $Z = d(V)$, and Ψ_n^* the final estimate of the data-generating distribution.

It is easy to see that $P_n D^*(d_n, \tau_n, P_n^*) = 0$: We have $P_n D_1(d_n, P_n^*) = P_n \frac{d}{d\varepsilon} L(Q_n(\varepsilon|g_n), g_n, (O_1, \ldots, O_n))|_{\varepsilon=0} = 0$; $P_n D_2(d_n, P_n^*) = 0$ when we are using the empirical distribution $P_{W,n}$; and $P_n D_3(d_n, \tau_n, P_n^*) = 0$ is described in the proof of optimality of the closed-form solution in Toth (2016).

1.6.3 Theoretical Results for Ψ_Z^*

1.6.3.1 Conditions for Efficiency of Ψ_Z^*

These six conditions are needed to prove asymptotic efficiency (Theorem 2). As discussed in Toth (2016), when all relevant components and nuisance parameters ($P_{W,n}$, ρ_n, μ_n) are consistent, then (C3) and (C4) hold, while (C6) holds by construction of the TMLE estimator.

(C1) $\rho_0(W)$ satisfies the strong positivity assumption: $Pr_0(\delta < \rho_0(W) < 1 - \delta) = 1$ for some $\delta > 0$.

(C2) The estimate $\rho_n(W)$ satisfies the strong positivity assumption, for a fixed $\delta > 0$ with probability approaching 1, so we have $Pr_0(\delta < \rho_n(W) < 1 - \delta) \to 1$.

Define second-order terms as follows:

$$R_1(d, P) \triangleq E_{P_0}\left[\left(1 - \frac{Pr_{P_0}(Z = d|W)}{Pr_P(Z = d|W)}\right)(\mu_P(Z = d, W) - \mu_0(Z = d, W))\right]$$

$$R_2(d, \tau_0, P) \triangleq E_{P_0}\Big[(d - d_0)(\mu_{b,0}(V) - \tau_0 c_{b,0}(V))\Big]$$

Let $R_0(d, \tau_0, P) = R_1(d, P) + R_2(d, \tau_0, P)$.

(C3) $R_0(d_n, \tau_0, P_n^*) = o_{P_0}(n^{-\frac{1}{2}})$.

(C4) $P_0[(D^*(d_n, \tau_0, P_n^*) - D_0)^2] = o_{P_0}(1)$.

C5) $D^*(d_n, \tau_0, P_n^*)$ belongs to a P_0-Donsker class with probability approaching 1.

(C6) $\frac{1}{n}\sum_{i=1}^n D^*(d_n, \tau_0, P_n^*)(O_i) = o_{P_0}(n^{-\frac{1}{2}})$.

1.6.3.2 Sufficient Conditions for Lemma 3

(E1) GC-like property for $c_{b,Z}(V)$, $\mu_{b,n}(V)$:
$\sup_V |(E_{W,n|V} - E_{W,0|V})c_{b,T}(W)| = \sup_V(|c_{b,Z,n}(V) - c_{b,Z,0}(V)|) = o_{P_0}(1)$

(E2) $\sup_V |E_{W,0|V}\mu_{b,n}(W) - E_{W,0|V}\mu_{b,0}(W)| = o_{P_0}(1)$

(E3) $S_n(x)$, defined as $x \to E_{V,n}[I(T_n(V) \geq x)c_{b,Z,n}(V)]$ is a GC-class.

(E4) Convergence of ρ_n, μ_n to ρ_0, μ_0, respectively, in $L^2(P_0)$ norm at a $O(n^{-1/2})$ rate in each case.

When all relevant components and nuisance parameters are consistent, as is the case when Theorem 2 below holds and our estimator is efficient, we also expect conditions (E1)–(E4) to hold.

Toth (2016) discusses the assumptions and conditions above in detail.

1.6.3.3 Efficiency and Inference

Theorem 2 (Ψ_Z^* is asymptotically linear and efficient.) *Assume assumptions (A1)–(A4) and conditions (C1)–(C6). Then, $\Psi_Z^* = \Psi_Z(P_n^*) = \Psi_{Z,d_n}(P_n^*)$ as defined by the TMLE procedure is a RAL estimator of $\Psi_Z(P_0)$ with influence curve D_0, so*

$$\Psi_Z(P_n^*) - \Psi_Z(P_0) = \frac{1}{n}\sum_{i=1}^n D_0(O_i) + o_{P_0}(n^{-\frac{1}{2}}).$$

Further, Ψ_Z^* is efficient among all RAL estimators of $\Psi_Z(P_0)$.

Inference. Let $\sigma_0^2 = Var_{O\sim P_0} D_0(O)$. By Theorem 2 and the central limit theorem, $\sqrt{n}(\Psi_Z(P_n^*) - \Psi_Z(P_0))$ converges in distribution to a $N(0, \sigma_0^2)$ distribution. Let $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n D^*(d_n, \tau_n, P_n^*)(O_i)^2$ be an estimate of σ_0^2.

Lemma 3 *Under the assumptions (C1) and (C2), and conditions (E1)–(E4), we have* $\sigma_n \to_{P_0} \sigma_0$. *Thus, an asymptotically valid 2-sided* $1-\alpha$ *confidence interval is given by*

$$\Psi_Z^* \pm z_{1-\frac{\alpha}{2}} \frac{\sigma_n}{\sqrt{n}}$$

where $z_{1-\frac{\alpha}{2}}$ denotes the $(1-\frac{\alpha}{2})$-quantile of a $N(0,1)$ r.v.

1.6.3.4 Double Robustness of $\Psi^*_{Z,n}$

Theorem 2 demonstrates consistency and efficiency when all relevant components and nuisance parameters are consistently estimated. Another important issue is under what cases of partial misspecification we still get a consistent estimate of $\Psi_{Z,0}$, albeit an inefficient one. Our TMLE-based estimate Ψ_Z^* is a consistent estimate of $\Psi_{Z,0}$ under misspecification of $\rho_n(W)$ in the initial estimates, but not under misspecification of $\mu_n(W, Z)$. However, it turns out there is still an important double robustness property. If we consider $\Psi_Z^* = \Psi_{Z,d_n}(P_n^*)$ as an estimate of $\Psi_{Z,d_n}(P_0)$, where the optimal decision rule $d_n(V)$ is estimated from the data, then we have that Ψ_Z^* is double robust to misspecification of ρ_n or μ_n in the initial estimates.

Lemma 4 (Ψ_Z^* is a double robust estimator of $\Psi_{Z,d_n}(P_0)$.) *Assume assumptions (A1)–(A4) and conditions (C1)–(C2). Also assume the following version of (C4):* $Var_{O\sim P_0}(D_1(d_n, P_n^*)(O) + D_2(d_n, P_n^*)(O)) < \infty$.
Then, $\Psi_Z^* = \Psi_{Z,d_n}(P_n^*)$ *is a consistent estimator of* $\Psi_{Z,d_n}(P_0)$ *when either* μ_n *is specified correctly, or* ρ_n *is specified correctly.*

The proof of this lemma is based on the equation

$$\Psi_{Z,d_n}(P_n^*) - \Psi_{Z,d_n}(P_0) = -P_0\big[D_1(d_n, P_n^*) + D_2(d_n, P_n^*)\big] + R_1(d_n, P_n^*)$$

where D_1, D_2, and R_1 are as defined in Sects. 1.6.1 and 1.6.3.1.

1.7 TMLE for Optimal Treatment Problem ($\Psi_{A,0}$)

We now present results for the case of intervening on the treatment, setting $A = d(V)$.

1.7.1 Efficient Influence Curve $D_A^*(\Psi_0)$

Lemma 5 *Let*

$$J_0(Z, W) = \frac{I(Z=1)}{\rho_0(W)} + \frac{\left(\frac{I(Z=1)}{\rho_0(W)} - \frac{I(Z=0)}{1-\rho_0(W)}\right)\left(d_0(V) - \Pi_0(W, Z=1)\right)}{\Pi_0(W, Z=1) - \Pi_0(W, Z=0)}$$

The efficient influence curve $D_A^*(\Psi_0)$ *is*

$$D_A^*(\Psi_0) = -\tau_0 E_{P_0}[c_T(d_0(V), W) - K] \tag{1.7}$$
$$+m_0(W)d_0(V) + \theta_0(W) - \Psi_0 \tag{1.8}$$
$$-J_0(Z, W)m_0(W)\left[A - \Pi_0(W, Z)\right] \tag{1.9}$$
$$+J_0(Z, W)\left[Y - (m_0(W)\Pi_0(W, Z) - \theta_0(W))\right] \tag{1.10}$$

We also write $D^*(d_0, \tau_0, P_0)$. For convenience, denote lines (1)–(4) of D^* above as D_c^*, D_W^*, D_Π^*, and D_μ^*, respectively. Finally, let D_{A,d_n}^* denote the efficient influence curve for $\Psi_{A,d_n}(P_0) \triangleq E_{P_{W,0}} m_0(W)d_n(V) + \theta_0(W)$, which is the mean counterfactual estimate when the decision rule is estimated from the data. We have $D_{A,d_n}^* = D_W^* + D_\Pi^* + D_\mu^*$ (see Toth 2016).

1.7.2 Iterative TMLE Estimator

We have derived two different TMLE-based estimators for $\Psi_{A,0}$. We present an iterative estimator here, which involves a standard, numerically well-behaved, and easily understood likelihood maximization operation at each step. The other estimator uses a logistic fluctuation in a single non-iterative step and has the advantage that the estimate μ respects the bounds of Y found in the data (see Toth 2016; Toth and van der Laan 2016).

The relevant components for estimating $\Psi_A = E_W[m(W)d(V) + \theta(W)]$ are $Q = (P_W, m, \theta)$. The nuisance parameters are $g = (\rho, \Pi)$. $d(V)$ and τ can be thought of as functions of P_W, m here. Let

$h_1(W) \triangleq \frac{1}{\rho(W)(\Pi(W,1)-\Pi(W,0))} + \frac{d(V)-\Pi(W,1)}{(\Pi(W,1)-\Pi(W,0))^2}\frac{1}{\rho(W)(1-\rho(W))}$. Also, let $h_2(W) \triangleq \frac{1}{\rho}\left[1 - \frac{\Pi(W,1)}{\Pi(W,1)-\Pi(W,0)} + \frac{d-\Pi(W,1)}{\Pi(W,1)-\Pi(W,0)}(1 - \frac{\Pi(W,1)}{\Pi(W,1)-\Pi(W,0)}\frac{1}{1-\rho})\right]$.
Then, we have that $D_\mu^* = (h_1\Pi + h_2)(Y - m\Pi - \theta)$.

If A is not binary, convert A to the unit interval via a linear transformation $A \to \tilde{A}$ so that $\tilde{A} = 0$ corresponds to $A_{\min}$ and $\tilde{A} = 1$ to $A_{\max}$. We assume $A \in [0, 1]$ from here.

1. Use the empirical distribution $P_{W,n}$ to estimate P_W. Make initial estimates of $Q = \{m_n(W), \theta_n(W)\}$ and $g_n = \{\rho_n(W), \Pi_n(W, Z)\}$ using any strategy desired. Data-adaptive learning using Super Learner is recommended.

2. The empirical estimate $P_{W,n}$ gives an estimate of $Pr_{V,n}(V) = E_{W,n}I(F_V(W) = V)$, $K_{A,n} = E_{W,n}c_A(0, W)$, $\overline{K_{A,n}} = E_{W,n}c_A(1, W)$, and $c_{b,A,n}(V) = E_{W,n|V}(c_A(\overline{1, W}) - c_A(0, W))$.
3. Estimate $m_n(V)$ as $E_{W,n|V}m(W)$.
4. Estimate $T_0(V)$ as $T_n(V) = \frac{m_n(V)}{c_{b,A,n}(V)}$.
5. Estimate $S_0(x)$ using $S_n(x) = E_{V,n}[I(T_n(V) \geq x)(c_{b,A,n}(V))]$.
6. Estimate η_0 as using $\eta_n = S_n^{-1}(K - K_{A,n})$ and $\tau_n = \max\{0, \eta_n\}$.
7. Estimate the decision rule as $d_n(V) = \overline{1}$ iff $T_n(V) \geq \tau_n$ (the decision rule is not updated iteratively).

ITERATE STEPS (8)–(9) UNTIL CONVERGENCE:

8. Fluctuate the initial estimate of $m_n(W)$, $\theta_n(W)$ as follows: Using $\mu_n(Z, W) = m_n(W)\Pi_n(Z, W) + \theta_n(W)$, run an OLS regression:
   Outcome: $(Y_i : i = 1, \ldots, n)$
   Offset: $(\mu_n(Z_i, W_i), i = 1, \ldots, n)$
   Covariate: $(h_1(W_i)\Pi_n(Z_i, W_i) + h_2(W_i) : i = 1, \ldots, n)$
   Let ε_n represent the level of fluctuation, with
   $\varepsilon_n = \text{argmax}_\varepsilon \frac{1}{n}\sum_{i=1}^n (Y_i - \mu_n(\varepsilon)(Z_i, W_i))^2$ and
   $\mu_n(\varepsilon)(Z, W) = \mu_n(Z, W) + \varepsilon(h_1(W)\Pi_n(Z, W) + h_2(W))$.

   Note that $\mu_n(\varepsilon) = (m_n + \varepsilon h_1)\Pi_n + (\theta_n + \varepsilon h_2)$ stays in the semiparametric regression model.
   Update m_n to $m_n(\varepsilon) = m_n + \varepsilon h_1$, θ_n to $\theta_n(\varepsilon) = \theta_n + \varepsilon h_2$.
9. Now fluctuate the initial estimate of $\Pi_n(Z, W)$ as follows: Use covariate $J(Z, W)$ as defined in Lemma 5. Run a logistic regression using:

   Outcome: $(A_i : i = 1, \ldots, n)$
   Offset: $(\text{logit}\, \Pi_n(Z_i, W_i), i = 1, \ldots, n)$
   Covariate: $(J(Z_i, W_i)m(W_i) : i = 1, \ldots, n)$
   Let ε_n represent the level of fluctuation, with
   $\varepsilon_n = \text{argmax}_\varepsilon \frac{1}{n}\sum_{i=1}^n [\Pi_n(\varepsilon)(Z_i, W_i) \log A_i + (1 - \Pi_n(\varepsilon)(Z_i, W_i))$
   $\log(1 - A_i)]$ and $\Pi_n(\varepsilon)(Z, W) = \text{logit}^{-1}(\text{logit}\, \Pi_n(Z, W) + \varepsilon J(Z, W)m(W))$.
   Update Π_n to $\Pi_n(\varepsilon)$. Also update $h_1(W)$, $h_2(W)$ to reflect the new Π_n.

---

10. Finally, form final estimate of $\Psi_{A,0} = \Psi_{A,d_0}(P_0)$, using a plug-in estimator with the final estimates upon convergence m_n^* and θ_n^*:

$$\Psi_A^* = \Psi_{A,d_n}(P_n^*) = \frac{1}{n}\sum_{i=1}^n \left[m_n^*(W_i) \cdot d_n(V_i) + \theta_n^*(W_i)\right]$$

As for Ψ_Z, it is straightforward to check that the efficient influence equation $P_n D^*(d_n, \tau_n, P_n^*) = 0$.

1.7.3 Double Robustness of Ψ_A^*

As in Sect. 1.6.3.4, Ψ_A^* is not a double robust estimator of $\Psi_{A,0}$: Component $m(W)$ must always be consistently specified as a necessary condition for consistency of Ψ_A^*. However, if we consider $\Psi_A^* = \Psi_{A,d_n}(P_n^*)$ as an estimate of $\Psi_{A,d_n}(P_0)$, where the optimal decision rule $d_n(V)$ is estimated from the data, then we have that Ψ_A^* is double robust:

Lemma 6 (Ψ_A^* is a double robust estimator of $\Psi_{A,d_n}(P_0)$.) *Assume (A1)–(A4) and (C1)–(C2). Also assume* $Var_{O\sim P_0}(D_d^*(d_n, P_n^*)(O)) < \infty$.
Then, $\Psi_A^* = \Psi_{A,d_n}(P_n^*)$ *is a consistent estimator of* $\Psi_{A,d_n}(P_0)$ *when either:*

- m_n *and* θ_n *are consistent*
- ρ_n *and* Π_n *are consistent*
- m_n *and* ρ_n *are consistent*

Above D_d^* refers to $D_\mu^* + D_\Pi^* + D_W^*$, the portions of the efficient influence curve that are orthogonal to variation in decision rule d. The proof is straightforward (see Toth 2016).

1.8 Simulations

1.8.1 Setup

We use two main data-generating functions:

Dataset 1 (categorical Y).

Data is generated according to:

$$
\begin{aligned}
U_{AY} &\sim \text{Bernoulli}(1/2)\\
W1 &\sim \text{Uniform}(-1, 1)\\
W2 &\sim \text{Bernoulli}(1/2)\\
Z &\sim \text{Bernoulli}(\alpha)\\
A &\sim \text{Bernoulli}(W1 + 10 \cdot Z + 2 \cdot U_{AY} - 10)\\
Y &\sim \text{Bernoulli}((1 - A) * (\text{plogis}(W2 - 2 - U_{A,Y})) + (A) * (\text{plogis}(W1 + 4))
\end{aligned}
$$

$U_{A,Y}$ is the confounding term. For the simulations where $V \subset W$, we take $V = (1(W1 \geq 0) + -1(W1 < 0), W2)$. We have $c_T(Z = 1, W) = 1$, $c_T(Z = 0, W) = 0$ for all W here.

Dataset 2 (continuous Y.)

We use three-dimensional W and distribution

$$\begin{aligned} U_{AY} &\sim \text{Normal}(0, 1) \\ W &\sim \text{Normal}(\mu_\beta, \Sigma) \\ Z &\sim \text{Bernoulli}(0.1) \\ A &\sim -2 \cdot W1 + W2^2 + 4 \cdot W3 \cdot Z + U_{AY} \\ Y &\sim 0.5 \cdot W1 \cdot W2 - W3 + 3 \cdot A \cdot W2 + U_{AY} \end{aligned}$$

When $V \subset W$, we use either V equals $W1$ rounded to the nearest 0.2, or alternately, V is $W3$ rounded to the nearest 0.2. We also have $c_T(0, W) = 0$ for all W, and $c_T(1, W) = 1 + b \cdot W1$, and varying μ_β, Σ, and b.

Forming initial estimates.
We use the empirical distribution $P_{W,n}$ for the distribution of W. For learning μ_n, we use Super Learner, with the following libraries of learners (the names of learners are as specified in the SuperLearner package (van der Laan et al. 2007):

For continuous Y: glm, step, randomForest, nnet, svm, polymars, rpart, ridge, glmnet, gam, bayesglm, loess, mean.

For categorical Y: glm, step, svm, step.interaction, glm.interaction, nnet.4, gam, randomForest, knn, mean, glmnet, rpart.

Further, we included different parameterizations of some of the learners given above, such as ntree = 100, 300, 500, 1000 for randomForest.

Finally, for learning ρ_n, we use a correctly specified logistic regression, regressing Z on W (except for simulation (C) as described below).

Estimators used.
For both parameters of interest Ψ_Z and Ψ_A, we report results on the TMLE estimator Ψ_Z^* (or Ψ_A^*), and the initial substitution estimator $\Psi_{Z,n}^0$ (or $\Psi_{A,n}^0$). The latter is the plug-in estimate, for instance $\Psi_{Z,n}^0 \triangleq \Psi_Z(P_{W,n}, \mu_n)$, that uses the same initial estimates of relevant components and the nuisance parameter as TMLE. Thus, the initial substitution estimator gives a comparison of TMLE to a straightforward semiparametric, machine learning-based approach. 1000 repetitions are done of each simulation.

Table 1.1 (Simulation A.) Consistent estimation of $\Psi_{Z,0}$ using machine learning, categorical $Y \cdot \Psi_{Z,0} = 0.3456$, $K = 0.3$, and $V \subset W \cdot \sigma_n^2 = \mathrm{Var}_{O\sim P_n} D_Z^*(d_n, \tau_n, P_n^*)(O)$

N = 250					
Estimator	Ψ_Z^*	Bias	Var	σ_n^2/N	Cover
TMLE	0.3545	0.0089	0.0071	0.0010	88.3
CV-TMLE	0.3541	0.0085	0.0017	0.0010	90.6
Init. Substit.	0.3427	−0.0029	0.0067	0.0010	(87.9)
N = 1000					
TMLE	0.3485	0.0029	0.0003	0.0003	93.3
CV-TMLE	0.3497	0.0041	0.0002	0.0003	96.8
Init. Substit.	0.3344	−0.0112	0.0003	0.0003	(88.3)
N = 4000					
TMLE	0.3467	0.0011	0.0001	0.0001	95.0
CV-TMLE	0.3498	0.0002	0.0001	0.0001	94.7
Init. Substit.	0.3429	−0.0027	0.0001	0.0001	(93.3)

Table 1.2 (Simulation B.) Consistent estimation of $\Psi_{A,0}$ using machine learning, continuous $Y \cdot \Psi_{A,0} = 336.2$, $K = 0.8$, and $V \subset W \cdot \sigma_n^2 = \mathrm{Var}_{O\sim P_n} D_Z^*(d_n, \tau_n, P_n^*)(O)$

N = 250					
Estimator	Ψ_A^*	Bias	Var	σ_n^2/N	Cover
TMLE	327.5	−8.7	344.7	176.3	78.4
Init. Substit.	310.0	−26.2	495.1	174.0	(47.8)
N = 1000					
TMLE	332.9	−3.3	40.7	38.5	89.0
Init. Substit.	322.7	−13.5	126.8	43.1	(53.2)
N = 4000					
TMLE	334.5	−1.7	8.4	9.1	93.3
Init. Substit.	328.7	−7.5	25.9	8.8	(41.3)

Simulations (A–B): using a large library of learning algorithms for consistent initial estimates.

Tables 1.1 and 1.2 show the behavior of our estimators when machine learning is used to consistently estimate all relevant components and nuisance parameters. Table 1.1 deals with estimating Ψ_Z when Y is categorical. In this case, bias is very low with or without the TMLE fluctuation step. σ_n^2/n gives a consistent estimate of the variance of Ψ_Z^*, in this case where efficiency holds. We see that both estimators have very low variance that converges to σ_n^2/n by $n = 1000$. Coverage of 95% confidence intervals is also displayed, with intervals calculated as $\Psi_n^* \pm 1.96\frac{\sigma_n}{\sqrt{n}}$, as in Lemma 3. The coverage is given in parentheses for the initial substitution estimator, as σ_n^2 is not necessarily the right variance. The TMLE estimators show better coverage,

even though, in this example, the width of the confidence intervals was accurate for all estimators for $n \geq 1000$. This may be due to the asymptotic linearity property of the TMLE-based estimators, ensuring that they follow a normal distribution as n becomes large.

Y is continuous in Table 1.2. TMLE convincingly outperforms the initial substitution estimator in both bias and variance here. Only the TMLE estimator is guaranteed to be efficient, and we see a significant improvement in variance. The estimated asymptotic variance σ_n^2/n approximates the variance seen in Ψ_A^* fairly well for $n \geq 1000$. The coverage of confidence intervals for TMLE seems to converge to 95% more slowly than for the previous case of categorical Y.

Simulation (C): double robustness under partial misspecification.
As described in Sect. 1.7.3, $\Psi_A^* = \Psi_{A,d_n}^*$ is a double robust estimator of $\Psi_{A,d_n}(\Psi_0)$, but not necessarily of $\Psi_{A,0}$.

Table 1.3 verifies consistency of Ψ_A^* when the initial estimate for μ_n is grossly misspecified as $\mu_n = \text{mean}(Y)$. This creates a discrepancy of $\sim$0.1 points between $\Psi_{A,d_n}(P_0)$ and $\Psi_{A,0}$. The initial substitution estimator retains a bias of around -0.09 in estimating $\Psi_{A,d_n}(P_0)$, while TMLE demonstrates practically zero bias by $n = 1000$. TMLE is not efficient in this setting of partial misspecification. It has significantly larger variance than the initial substitution estimator for smaller sample sizes, but the variances are similar by $n = 4000$. For confidence intervals, the width was calculated by estimating $\text{Var}(\Psi_{A,d_n})$ as $\sigma_n^2 = \text{Var}_{O\sim P_n} D_{d_n}^*(P_n^*)(O)$, where $D_{d_n}^*(P)$ is the effi-

Table 1.3 (Simulation C.) Robustness of Ψ_A^* to partial misspecification, μ_n is misspecified. $\Psi_{A,0} = 0.63$, $K = 0.5$, and $V = W$

N = 1000					
Estimator	Ψ_A^*	$\left(\Psi^* - \Psi_{d_n}(P_0)\right)$	$\left(\Psi^* - \Psi_0\right)$	Var	Cover
TMLE	0.54	0.00	−0.09	0.69	93.3
Init. Substit.	0.45	−0.10	−0.18	0.24	(69.2)
N = 4000					
TMLE	0.54	0.00	−0.09	0.11	96.8
Init. Substit.	0.45	−0.09	−0.18	0.10	(40.1)

Table 1.4 (Simulation D.) Estimation of true mean outcome $\Psi_{Z,d_n}(P_0)$, under rule $d_n \cdot \Psi_{Z,d_n}(P_0) = 162.8$ when $K = 0.2$, and $\Psi_{Z,d_n}(P_0) = 289.1$ when $K = 0.8$. Sample size is $N = 1000$ and $V = W$

	K = 0.2		K = 0.8	
Learning μ_n	Ψ_Z^*	Var	Ψ_Z^*	Var
Large library	158.9	8.14	286.4	9.32
Small library	148.3	49.45	267.9	16.28
No fitting	142.2	12.83	264.1	10.30

cient influence curve of $\Psi_{A,d_n}(P)$ as defined in Sect. 1.7.1. It provides a conservative (over)-estimate of variance for confidence intervals, as discussed in Toth and van der Laan (2016). We see that TMLE's coverage converges to just above 95%. On the other hand, coverage is very low for the initial substitution estimator due to its bias. This is despite the fact that the intervals are too wide in this case.

Simulation (D): quality of the estimate of d_n versus the true mean outcome attained under rule d_n.
We study how more accurate estimation of the decision rule d_n can lead to a higher objective obtained. The objective maximized here is the mean outcome under rule d_n, where d_n must satisfy a cost constraint. We use the known true distributions for $P_{W,0}$ and μ_0 in calculating the value of mean outcome under d_n as $\Psi_{d_n}(P_0) = E_{P_0}\mu_0(W, Z = d_n(V))$. The highest the true mean outcome can be under a decision rule that satisfies $E_{P_0}c_T(W, Z = d(V)) \leq K$ is Ψ_0 using optimal rule $d = d_0$. Therefore, the discrepancy between $\Psi_{d_n}(P_0)$ and Ψ_0 gives a measure of how inaccurate estimation of the decision rule diminishes the objective.

We compare $\Psi_{d_n}(P_0)$ when estimating μ_n using the usual large library of learners; when using a smaller library of learners consisting of mean, loess, nnet.size = 3, nnet.size = 4, nnet.size = 5; and finally when we set $\mu_n = \text{mean}(Y) \cdot d_n$ is estimated using μ_n as usual (note that it is the same between the initial substitution and TMLE-based estimates). Table 1.4 confirms the importance of forming a good fit with the data for achieving a high mean outcome. For $K = 0.2$ when roughly 20% of the population could be assigned $Z = 1$, the mean outcome was only a few points below the true optimal mean outcome Ψ_{d_0}, when using the full library of learners (158.9 vs. 162.8). However, it was about 15 points lower when using a much smaller library of learners. In fact, even when using machine learning with several nonparametric methods in the case of the smaller library, the objective $\Psi_{d_n}(P_0)$ attained was not far from that attained with the most uninformative $\mu_n = \text{mean}(Y)$. Very similar results hold for the less constrained case of $K = 0.8$.

1.9 Discussion

We considered the resource-allocation problem of finding the optimal mean counterfactual outcome given a general cost constraint, in the setting where unmeasured confounding is a possibility and an instrumental variable is available. This work dealt with both problems of finding an optimal treatment regime, and finding the optimal intent-to-treat regime. For both cases, we gave closed-form solutions of the optimal intervention and derived estimators for the optimal mean counterfactual outcome. Our model allows the individualized treatment (or intent-to-treat) rules to be a function of an arbitrary subset of baseline covariates. Estimation is done using the targeted maximum likelihood (TMLE) methodology, which is a semiparametric approach having a number of desirable properties (efficiency, robustness to misspecification, asymptotic normality, and being a substitution estimator). Simulation

results showed that TMLE can simultaneously demonstrate both finite-sample bias reduction and lower variance than straightforward machine learning approaches. The empirical variance of TMLE estimators appears to converge to the semiparametric efficiency bound, and confidence intervals are accurate for sample sizes of a few thousand. Consistency in the case of partial misspecification was confirmed, in the sense of Lemmas 4 and 6. Our simulations also addressed the important question of to what extent improved statistical estimation can lead to better optimization results. We were able to demonstrate significant increases in the value of the mean outcome under the estimated optimal rule, when a larger library of data-adaptive learners achieved a closer fit.

References

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–471.

Brookhart, M. A., & Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects. *International Journal of Biostatistics*, *3*(1), 1–14.

Brookhart, M. A., Rassen, J. A., & Schneeweiss, S. (2010). Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiology and Drug Safety*, *19*(6), 537–554.

Chakraborty, B., & Moodie, E. E. (2013). *Statistical methods for dynamic treatment regimes*. Berlin Heidelberg New York: Springer.

Chakraborty, B., Laber, E., & Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, *69*(3), 714–723.

Chesney, M. A. (2006). The elusive gold standard. Future perspectives for HIV adherence assessment and intervention. *Journal of Acquired Immune Deficiency Syndromes*, *43*(1), S149–155.

Editors: National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. (2011). Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease. National Academies Press (US), Washington DC.

Gruber, S., & van der Laan, M. (2010). A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *International Journal of Biostatistics*, **6**(1). Article 26.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*, 467–475.

Karp, R. (1972). *Reducibility among combinatorial problems*. New York Berlin Heidelberg: Springer.

Luedtke, A., & van der Laan, M. (2016a). Optimal individualized treatments in resource-limited settings. *International Journal of Biostatistics*, *12*(1), 283–303.

Luedtke, A., & van der Laan, M. (2016b). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics*, *44*(2), 713–742.

Newey, W. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, *5*(2), 99–135.

Pearl, J. (2000). Causality: Models, reasoning, and inference. Cambridge University Press.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. in *Proceeding of the Second Seattle Symposium in Biostatistics* (Vol. 179, pp. 189–326).

Toth, B. (2016). Targeted learning of individual effects and individualized treatments using an instrumental variable. PhD dissertation, U.C. Berkeley.

Toth, B., & van der Laan, M. (2016). TMLE for marginal structural models based on an instrument. U.C. Berkeley Division of Biostatistics Working Papers Series, working paper 350.

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

van der Laan, M., Rubin, D. (2006). Targeted maximum likelihood learning. *International Journal of Biostatistics*, *2*(1). Article 11.

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

Zhang, B., Tsiatis, A., Davidian, M., Zhang, M., & Laber, E. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, *68*, 1010–1018.

Chapter 2
Overview of Omics Biomarker Discovery and Design Considerations for Biomarker-Informed Clinical Trials

Weidong Zhang, Bo Huang, Jing Wang and Sandeep Menon

2.1 Overview of Omics Biomarker Discovery

Biomarker discovery is critical in drug development to understand disease etiology and evaluate drug activity. Rapid development of omics technologies over the last few decades has offered tremendous opportunities in biomarker discovery. Omics biomarkers are high dimensional in nature. The major omics technologies include genomics (the study of genes, mutations and their functions), transcriptomics (the study of the mRNA and their expressions), proteomics (the study of proteins and their expressions), metabolomics (the study of molecules involved in cellular metabolism), lipomics (the study of cellular lipids and their functions), and glycomics (the study of cellular carbohydrates and their functions). Omics biomarkers are high dimensional in nature. Each omics technology may output thousands or millions of analytes, and dimensionality of omics data varies from a few hundreds to a few millions. Advancements in omics technologies have provided us great opportunities to understand disease biology from the unbiased global landscape. The technology boom started from late twentieth century when Microarray was first available for measurement

W. Zhang (✉) · J. Wang
Global Product Development, Pfizer Inc., 3rd FL, 300 Technology Square, Cambridge, MA 02139, USA
e-mail: weidong.zhang2@pfizer.com

J. Wang
e-mail: Jing.Wang122@pfizer.com

B. Huang
Global Product Development, Pfizer Inc., 280 Shennecossett Rd, Groton, CT 06340, USA
e-mail: Bo.Huang@pfizer.com

S. Menon
Worldwide Research and Development, Pfizer Inc., 1 Portland Street, Cambridge, MA 02139, USA
e-mail: Sandeep.M.Menon@pfizer.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_2

of whole transcriptome and genome. A DNA microarray is a solid surface attached with a collection of DNA fragments, known as probes or oligos. A probe is a specific sequence of a section of a gene that can be used to hybridize a cDNA or cRNA from a fluorescent molecule-labeled target sample. The fluorescent intensity of a probe—target hybridization is measured and quantified to determine the abundance of DNA molecules in the target sample. As a novel alternative strategy for genomics study, next-generation sequencing (NGS) technologies emerged after the completion of the Human Genome Project in 2003 have completely revolutionized biomedical research in the last decade. As a result, both turnaround time and cost of sequencing have been substantially reduced. According to the National Human Genome Research Institute (NHGRI), the cost of sequencing a genome dropped from $100 million in 2001 to $1245 in 2015 (Wetterstrand 2016), and the turnaround time was shortened from years in the late 90s to days including analysis in 2016 (Meienberg et al. 2016). Over the last decade, NGS technology has been widely applied to biomedical research in variety of ways including transcriptome profiling, identification of new RNA splice variant, genome-wide genetic variants identification, genome-wide epigenetic modification, and DNA methylation profiling. NGS technology is particularly a good fit to cancer research given the "disorder of genome" nature of cancer disease. In cancer research, NGS has significantly enhanced our ability to conduct comprehensive characterization of cancer genome to identify novel genetic alterations and has significantly helped dissect tumor complexity. Coupling with sophisticated computational tools and algorithms, significant achievements have been accomplished for breast cancer, ovarian cancer, colorectal cancer, lung cancer, liver cancer, kidney cancer, head and neck cancer, melanoma, acute myeloid leukemia (AML), etc. (Shyr and Liu 2013).

2.2 Statistical Considerations in Omics Precision Medicine Study

2.2.1 Data Integration

Human diseases are mostly complex diseases that involve multiple biological components. Rapid rate of discovery has revealed many molecular biomarkers including omic biomarkers that are associated with disease phenotypes. However, translating those associations into disease mechanisms and applying the discovery to clinic remain a great challenge. In genome-wide association studies, the major issues are either the effects of associated variants are too small effects or those effects do not appear to be functionally relevant. For example, many genetic variants that may be responsible for certain genetic disposition of certain disease reside on non-coding region of the genome (Lowe and Reddy 2015). Using information from single biological process, e.g., genetic polymorphism, may limit our ability to unveil true biological mechanism. On the other hand, single data set from on experiment repre-

sents only one snapshot of the biology, which necessitates integration of data from multiple experiments or multiple biological processes. It is well known that precision medicine is a systems biology that requires a holistic approach to understand disease etiology. By nature, multiple data sets generated from different sources such as different laboratories, different omics data types may be studied together to achieve maximum benefits. However, integration of data from different technologies or platforms has posed steep challenges on data analysis (Bersanelli et al. 2016). A few statistical methods have been proposed but areas have been focused on multivariate analysis approaches such as Partial least squares (PLS), Principal Component Analysis (PCA), and network analysis (Bersanelli et al. 2016). A recent method developed by Pineda et al. seemed to work well on combining information from genetic variant, DNA methylation, and gene expression data measured in bladder tumor samples, in which study penalized regression methods (LASSO and Elastic NET), were employed to explore relationships between genetic variants, DNA methylation, and gene expression measured in bladder tumor samples (Pineda et al. 2015).

Another issue with data integration is data preprocessing and normalization. For example, one may want to combine gene expression data derived from platforms such as PCR, Microarray, or NGS. Or one may want to combine gene expression data from the same technology but the data are generated from different laboratories. To address these issues and ensure valid comparisons between data sets, cross-platform normalization has been proposed before data integration (Shabalin et al. 2008; Thompson et al. 2016).

2.2.2 Power and Sample Size Assessment

Power and sample size estimation in precision medicine using omics technology remain statistically challenging due to the high dimensionality and uncertain effect sizes. Numerous methods have been proposed for expression-based omics data. For example, Jung and Young developed a method to take the advantage of pilot data for confirmatory experiment controlling family-wise error rate (FWER). When pilot data are not available, a two-stage sample size recalculation was proposed using the first stage data as pilot data (Jung and Young 2012). A false discovery rate (FDR)-based approach for RNAseq experiment was developed by Bi and Liu, by which the average power across the differentially expressed genes was first calculated, and then a sample size to achieve a desired average power while controlling FDR was followed (Bi and Liu 2016). Similar FDR-based approach was also available for microarray or proteomics experiment (Liu and Hwang 2007).

Most power and sample size calculations focus on univariate analysis. However, there are growing needs to tackle this problem in multivariate analysis. Saccenti and Timmerman have proposed a method for sample size estimation in a multivariate principal component analysis (PCA) and partial least-squares discriminant analysis (PLS-DA) (Saccenti and Timmerman 2016). In the case of PCA, one may want

to determine minimal sample size to obtain stable and reproducible PCA loading estimates. For PLS-DA, the goal is to assess how sample size and variability of the data affect the sensitivity and specificity of classification using PLS-DA (Saccenti and Timmerman 2016). Although those algorithms were developed using certain omics data, each one of them may be used as a generalized approach to the data that have similar data type such as gene expression and metabolomics.

Sample size estimation in genome-wide association studies (GWAS) requires special treatment given its unique features as compared to other omics data, e.g., transcriptomics or proteomics data. Often GWAS is conducted in a case-control design or family-based (case-parents trio) design. Since GWAS typically evaluates hundreds of thousands of SNP markers, a much larger sample size is expected to achieve reasonable power (Klein 2007; Spencer et al. 2009; Wu and Zhao 2009; Park et al. 2010). The power and sample size depend on multiple factors including effect size, number of SNPs being tested, distribution of minor-allele frequency (MAF), disease prevalence, linkage disequilibrium (LD), case/control ratio, and assumption of error rate in an allelic test (Hong and Park 2012). Considering complexity of genetic study and data structure and objectives of the studies, numerous methods for sample size calculation have been proposed according to the specific scenarios (Jiang and Yu 2016; Lee et al. 2012).

2.2.3 Statistical Modeling

Conventional statistics focus on problems with large number of experimental units (n) as compared to small number of features or variables (p) measured from each unit. In drug discovery, biomarker discovery using omics data in precision medicine often deals with "large p, small n" problem, in which hundreds of thousands analytes are measured from relatively much smaller number of subjects (sometimes as few as a dozen). An array of statistical methods have been developed in analysis of high-dimensional omics data. Those methods include exploratory clustering analysis to investigate patterns and structures, and univariate or multivariate regression and classification analysis to predict disease status (Johnstone and Titterington 2009). For expression-based omics data such as gene expression, proteomics, metabolomics, dimension reduction is considered as the first step before subsequent analysis. Dimension reduction techniques include descriptive statistical approach such as coefficient variation (CV) filtering, by which analytes with low CV are removed from subsequent regression/ANOVA analysis. This approach was particularly useful when computing power is limited. Given today's high computing capacity, CV step is typically skipped and a univariate regression analysis is used for both dimension reduction and inference.

Although univariate single-analyte analysis is still a common approach for high-dimensional data due to its simplicity and interpretation benefit, multivariate and multiple regressions considering multiple analytes in a model become more popular for the advantage of: (1) Complexity of disease mechanism requires an inte-

grated information from multiple biomarkers to explain more biological variations. (2) Interactions between biomarkers cannot be modeled with single biomarker analysis. (3) Correlation and dependency among biomarkers cannot be handled with single-analyte analysis. Common multivariate methods include elastic net regularized regression (ENET), random forest (RF) and classification and regression trees (CART).

High-dimensional omics data are complex in regard to not only their dimensionality but their correlation structures. Therefore, controlling false discovery in high-dimensional omics data may need more statistical rigor. Family-wise error rate (FWER) adjustment techniques such as Bonferroni correction are easy to implement but generally considered too conservative. Benjamini and Hochberg (BH) introduced a sequential p-value procedure that controls FDR (Benjamini and Hochberg 1995). The BH method first finds the largest k such that $P_{(m)} \leq k/m * \alpha$, where m stands for m tests and α is a predefined FDR level, and rejects the null hypothesis for all H(i) for $I = 1 \ldots k$. Compared with FWER approach, BH is able to gain more power regarding statistical discoveries. Another FDR-related method, which is widely applied in the omics data analysis, is the q-value method developed by Storey (2002). The q-value is a measure of significance in terms of the false discovery rate. Both q-value and BH methods allow dependence of testing.

For GWAS, selection of genome-wide significance threshold is challenging due to the ultra-high number of statistical testing and complex genetic LD structures. Different procedures including Bonferroni, FDR, Sidak, permutation have been proposed; however, it was suggested that a $p = 5 \times 10^{-8}$ can be used for genome-wide significance and $p = 1 \times 10^{-7}$ can be used as a suggestive threshold at practical level (Panagiotou and Ioannidis 2012; Pe'er et al. 2008). A recent study from Fadista et al. further updated the thresholds by investigating different scenarios. They suggested that P-value thresholds should take into account impact of LD thresholds, MAF, and ancestry characteristics. A p-value threshold of 5×10^{-8} was confirmed for European population with MAF > 5%. However, the P-value threshold needs to be more stringent with European ancestry with low MAF (3×10^{-8} for MAF $\geq 1\%$, 2×10^{-8} for MAF $= 0.5\%$ and 1×10^{-8} for MAF $\geq 0.1\%$ at LD r2 < 0.8) (Fadista et al. 2016).

2.3 Biomarker-Informed Design Considerations

2.3.1 Classical Designs

Classical designs are widely used in clinical development of personalized medicine with a predictive biomarker which does not involve any pre-specified statistical adaptations based on the interim outcomes. Classical population enrichment designs can be categorized as two types of designs: retrospective enrichment design and prospective enrichment design.

2.3.1.1 Retrospective Designs

When prospective validation and testing of a biomarker is not feasible or not assessable in time at the beginning of the trial, retrospective enrichment design—a traditional all-comers design with retrospective validation of a biomarker, could be considered.

Retrospective validation is conducted after the completion of the study and may involve previously conducted trials in the same patient population. As stated by Mandrekar and Sargent (2009) when conducted appropriately, this design can aid in bringing forward effective treatments to marker-defined patient populations in a timely manner that might otherwise be impossible due to ethical and logistical (i.e., large trial and long time to complete) considerations. For such retrospective analysis to be valid and to minimize bias, Mandrekar and Sargent summarized a list of essential elements that are critical to retrospective validation studies.

- Data from a well-conducted randomized controlled trial
- Availability of samples on a large majority of patients to avoid selection bias
- Prospectively stated hypothesis, analysis techniques, and patient population
- Predefined and standardized assay and scoring system
- Upfront sample size and power justification.

Taking the development of EGFR-inhibitors cetuximab and panitumumab in metastatic colorectal cancer (CRC) as an example, Cetuximab and panitumumab were initially marketed for the indication of EGFR+CRC, which represents 65% of advanced colorectal cancer patients. Based on retrospective analysis of previously conducted randomized phase III and II trials (Karapetis et al. 2008; Bokemeyer et al. 2009; Van Cutsem et al. 2009), it has been demonstrated that cetuximab significantly improves the overall survival for patients with wild-type KRAS (a protein that in humans is encoded by the KRAS gene), with no survival benefit in patients harboring KRAS-mutant status. As a result, in July 2009, the FDA approved cetuximab for treatment of KRAS wild-type colon cancer. Similarly, in a prospectively specified analysis of data from a previous randomized phase III trial of panitumumab versus best supportive care (Amado et al. 2008), the hazard ratio for progression-free survival (PFS) comparing panitumumab with best supportive care in the KRAS wild-type and mutant subgroups was 0.45 and 0.99, respectively, with a statistically significant treatment by KRAS status interaction ($p < 0.0001$). Given the lack of activity in KRAS-mutant group, the label was changed to include wild-type patients only in 2009.

2.3.1.2 Prospective Designs

In contrast to retrospective enrichment designs that test and assess biomarker of interest retrospectively, prospective enrichment designs prospectively test, assess biomarkers, and select patients at the beginning of the trial. Although retrospective evaluation of predictive biomarkers could save resources and time and make effective

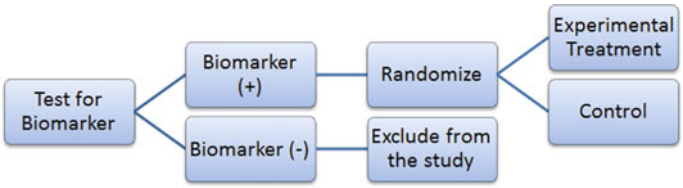


Fig. 2.1 Enrichment design

treatments available to patients in a much expedited timeframe, it may introduce serious bias due to the nature of retrospective selection of patient subgroups and lack of controlled validation of biomarkers. Hence, in the clinical development of targeted therapies and predictive biomarkers, prospective design is still the gold standard.

There are a number of prospective enrichment designs including the classical enrichment/targeted design, biomarker stratified design, sequential testing strategy design, biomarker-analysis design, marker-based strategy design and hybrid design. These designs differ from each other by the primary hypothesis test, randomization, multiplicity approaches which in turn affects the operating characteristics of the design including sample size, power, and type I error rate.

Enrichment/Targeted Design

In an enrichment design or targeted design, all patients in the trial may not generally benefit from the study treatment under consideration. The goal of the enrichment designs is to study the clinical benefit in a subgroup of the patient population defined by a specific biomarker status. In this design, the patients are screened for the presence or absence of a biomarker(s) profile. After extensive screening, only patients with the presence of a certain biomarker characteristic or profile are enrolled in the clinical trial (Freidlin et al. 2010; Sargent et al. 2005). In principle, this design essentially consists of an additional criterion for patient inclusion in the trial (Fig. 2.1).

A recent example for the enrichment design was of mutated *BRAF* kinase (Chapman et al. 2011). Almost 50% of melanomas have an activating V600E *BRAF* mutation. This leads to the hypothesis that inhibition of mutated *BRAF* kinase will have meaningful clinical benefit. Hence, only patients who tested positive for V600E *BRAF* mutation were enrolled in the study. Patients were randomized to an inhibitor of mutated *BRAF* kinase or control treatment. As hypothesized, the large treatment benefit was observed in the pre-specified subgroup. Another example for enriched design was used in HER2 trial where patients with HER2+ breast cancer (Romond et al. 2005). During the conduct of the study, it is important to have rapid turnaround times for the assay results in order to enroll patients faster. In addition, the assay testing should be consistent between different laboratories. For example, a high discordance was found between the local and central testing for HER2 status.

Hence, the Herceptin therapy may have benefitted a potentially larger group than defined as HER2+ by central testing.

The following considerations should be taken into account in this design—(1) a smaller sample size is usually required but the screening may still take the same amount of time (or even longer as explained below) as with an all-comers designs given the extensive pre-screen testing that will be conducted before enrollment; (2) the marketing label will be restricted; (3) there may still be a potential subset of patients who may benefit with the new treatment; (4) restricted enrollment does not provide data to establish that treatment is ineffective in biomarker negative patients; (5) a low prevalence of the marker may be challenging operationally and financially. Operationally, the biggest challenge is in recruitment, and financially, it may not be commercially attractive.

The efficiency of this design is a function of the percent of biomarker positive patients who are likely to be benefitted by the target treatment, and the reliability and reproducibility of the assay also play a pivotal role. This design is appropriate when the mechanistic behavior of drug is known and there is compelling preliminary evidence of benefit to a subset population.

Biomarker Stratified Design

It is also called as the biomarker by treatment interaction design. This design is most appropriate when there is no preliminary evidence to strongly favor restricting the trial to patients with specific biomarker profile that would necessitate a biomarker-enrichment design. This design is prospective and leads to a definitive marker validation strategy. In such cases, marker by treatment or stratified design is more informative than biomarker-enrichment design. In this design, the patients are tested for biomarker status and then separately randomized according to their positive or negative status of the marker (Freidlin et al. 2010). Thus, the randomization is done using marker status as the stratification factor; however, only the patients with a valid measurable marker result are randomized. Patients in each marker group are then randomized to two separate treatments (Fig. 2.2). Two separate hypotheses tests are conducted to determine the superiority of one treatment over the other separately within each marker group. The sample size is calculated separately to power the testing within each marker subgroup. Another variation to the hypothesis test within the same design is to conduct a formal marker by treatment interaction test to see if the treatment effect varies within each marker status subgroup. In this case, the study is powered based on the magnitude of interaction. This design can be viewed as two stand-alone trials; however, it is different from a large clinical trial by the calculation of the sample size and restriction of the randomization to only patients with a valid marker result.

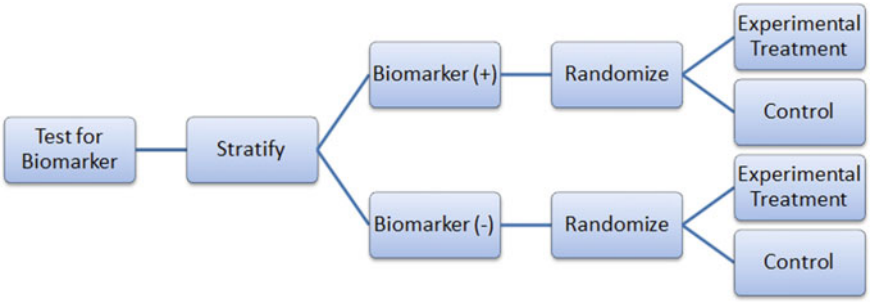


Fig. 2.2 Marker by treatment Interaction design

Sequential Testing Strategy Design

Sequential testing designs in principle can be considered as a special case of the classical randomized clinical trial for all comers or unselected patients.

a. Test Overall Difference Followed by Subgroup

Simon and Wang (2006) proposed an analysis strategy where the overall hypothesis is tested to see if there is a difference in the response in new treatment versus the control group. If there is no difference in the response is not significant at a pre-specified significance level (for example 0.01), then the new treatment is compared to the control group in the biomarker status positive patients. The second comparison uses a threshold of significance which is proportion of the traditional 0.05 not utilized by the initial test. This approach is useful when the new treatment is believed to be effective in a wider population, and the subset analysis is supplementary and used as a fallback option.

For example, if the trial is planned for having 90% power for detecting a uniform 33% reduction in overall hazard using a two-sided significance level of 0.03, then the overall analysis will take place when there are 297 events. If the test is positive in 25% of patients and the test is not prognostic, then at the time of analysis there will be approximately 75 events among the test positive patients. If the overall test of treatment effect is not significant, then the subset test will have 75% power for detecting a 50% reduction in hazard at a two-sided 0.02 significance level. By delaying the treatment evaluation in the test positive patients, 80% power can be achieved when there are 84 events and 90% power can be achieved when there are 109 events in the test positive subset.

Song and Chi (2007) later proposed a modification of the above method. Their method takes into account the correlation between the test statistics of the hypotheses of the overall population and the biomarker positive population.

b. Test Subgroup Followed by the Overall Population

Bauer (1991) investigated the multiple testing in the sequential sampling. Here, the hypothesis for the treatment is first tested in the biomarker positive status patients

and then tested in the overall population. This strategy is appropriate when there is a preliminary biological basis to believe that biomarker positive patients will benefit more from the new drug and there is sufficient marker prevalence to appropriately power the trial. In this closed testing procedure, the final type I error rate is always preserved. The approach of first testing in the subgroup defined by marker status has been implemented in the ongoing US-based phase III trial testing cetuximab in addition to infusional fluorouracil, leucovorin, and oxaliplatin as adjuvant therapy in stage III colon cancer (Albert et al. 2005). While the trial has now been amended to accrue only patients with *KRAS*–wild-type tumors, approximately 800 patients with *KRAS* mutant tumors have already been enrolled. In this trial, the primary analysis would be conducted at the 0.05 level in the patients with wild-type *KRAS*. A sample size of 1,035 patients with wild-type *KRAS* per arm would result in 515 total events, providing 90% power to detect a hazard ratio of 0.75 for this comparison using a two-sided log-rank test at a significance level of 0.05. If this subset analysis is statistically significant at $P=0.05$, then the efficacy of the regimen in the entire population will also be tested at level 0.05, as this is a closed testing procedure. This comparison using all 2910 patients will have 90% power to detect a hazard ratio of 0.79 comparing the two treatment arms, based on a total of 735 events.

Biomarker-Analysis Design

Biomarker-analysis design (Baker et al. 2012) essentially has two elements. The first element is a randomized trial with the presence or absence of the biomarker examined in all participants followed by the identification of a promising subgroup. This is done by using a plot of treatment benefit versus various cut-points or intervals of the biomarker. It is critical in this design that the specimen be at least collected at the randomization even if it is not examined, though it is preferred that it is examined at randomization. Collection of the specimens a priori can mitigate the risk of non-compliance with the treatment due to the incoming knowledge of the marker data (Baker and Freedman 1995). As the data trickles in, the investigators need to assess the risk and benefit especially from an ethical perspective of concealing the new information.

This design can assist in understanding the following hypothesis test: (a) targeted treatment versus standard of care in the overall population; (b) targeted treatment versus standard of care in the biomarker positive population; (c) targeted treatment versus standard of care in the biomarker negative population; (d) marker-based treatment selection versus targeted therapy, and (e) marker-based treatment selection versus standard of care. With multiple hypotheses that can be tested for (a) to (e), the significance levels and confidence intervals need to be adjusted according to the type and the number of hypotheses under consideration.

The selection of biomarker subgroup using the cut-points can be done using graphics. Various graphics have been proposed in the literature which assists in better visualization and understanding of the cut-points and intervals. Some of these plots present confidence intervals that adjust for multiplicity. Commonly used plots

include (a) marker-by-treatment predictiveness curves, (b) selection impact curve, (c) tail-oriented subgroup plot, and (d) the sliding-window subgroup plot.

(a) *Marker-by-treatment predictiveness curves* (Janes et al. 2011)—the risk of the response is plotted separately under targeted therapy and standard of care treatments for subjects with a marker in the interval.
(b) *Selection impact curve* (Song and Pepe 2004)—benefit of marker-based treatment selection is plotted directly as a function of marker cut-points.
(c) *Tail-oriented subgroup plot* (Bonetti and Gelber 2000)—The estimated benefit of targeted therapy against the standard of care is plotted among subjects with a marker level greater than a cut-point as a function of different clinically meaningful cut-points. Hence, the tail of the distribution is specified when the estimated benefit is plotted for a marker level above a certain cut-point.
(d) *Sliding-window subgroup plot* (Bonetti and Gelber 2004)—The estimated benefit of targeted therapy against the standard of care is plotted among subjects with a marker level within an interval as a function of marker level. Hence, the sliding window is specified when the estimated benefit is plotted for a marker level within a certain interval.

The tail-oriented subgroup plot and sliding window plots (Cai et al. 2011) give confidence intervals that account for multiple testing of several cut-points or intervals.

Baker and Kramer (2005) proposed a special case of biomarker analysis design for rare events. In this design, all subjects are randomized to either the targeted therapy or the control. The specimens are collected but not examined at the time of randomization. Subjects are randomly selected at the end of the trial to test for the presence or absence of the marker. The probability of random selection is ascertained based on the positive outcome of interest. King et al. (2001) proposed testing for the marker only for the subjects with a positive outcome of interest. This type of design can be referred to as biomarker-nested case-control design.

Marker-Based Strategy Design

In this design, patients are randomly assigned to treatment dependent or independent of the marker status (Fig. 2.3). All patients randomized to the non-biomarker-based arm receive the control treatment. In the biomarker-based arm, the patients receive the targeted or experimental therapy if the marker is positive and control treatment if the marker is negative (Freidlin et al. 2010; Sargent et al. 2005). The outcome of all of the patients in the marker-based subgroup is compared to that of all patients in the non-marker-based subgroup to investigate the predictive value of the marker. One downside of this design is that patients treated with the same regimen are included in both the marker-based and the non-marker-based subgroup, resulting in a substantial redundancy leading to many patients receiving the same treatment regimes in both subgroups. Hence, this design can reduce the treatment effect especially if the prevalence of the marker is high requiring a large sample size. This is illustrated in the following example, in the ERCC1 trial (Cobo et al. 2007) and presented in (Freidlin

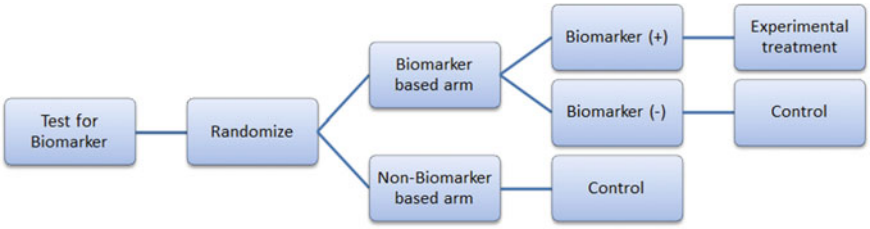


Fig. 2.3 Marker-based strategy design

et al. 2010) about 57% of the biomarker-based strategy arm patients were assigned to the same regimen of cisplatin + docetaxel as done in the standard of care arm. Thus, the comparison weakens the between-arm treatment effect difference and reduces the statistical power to reject null hypotheses. This can lead to either getting incomplete information or may miss a valuable biomarker in addition to delaying the evaluation of the biomarker due to the increased sample size required to achieve a desired power. One other disadvantage of this design is the inability to examine the effect of targeted therapy in patients in the negative marker status group as none of these patients receive it. Even if the patients in the negative marker status group respond to the targeted therapy, this cannot be assessed. The treatment difference between the new treatment and the control treatment can be diluted by marker-based treatment selection and sometimes can be a poor choice as compared to the randomized design.

A modified version of this design has been proposed where negative marker status group undergoes randomization and receives the targeted therapy or the control. Thus, the modified design allows assessment of the targeted therapy in both the biomarker positive and negative subgroup. It helps to assess whether the efficacy of the marker positive patients to therapy is because of the marker status being positive or due to an improved treatment regardless of the marker status. The assessment can also be done retrospectively if the classification of the marker needs to be revisited.

Hybrid Design

This design should be considered when there is strong evidence from preclinical or prior studies that there is efficacy of some treatment(s) on the marker-based subgroup. This makes it almost impossible due to ethical reasons to randomize patients with a particular marker to other treatment options. It is very similar to the enrichment designs, and all patients are examined for marker status and are randomly assigned to treatment or assigned to standard-of-care treatment for patients with positive biomarker values. However, only a marker-positive subgroup of patients is randomly assigned to treatments, whereas patients in the marker-negative group are assigned to control or standard-of-care treatment (Fig. 2.4). The study is powered to detect treatment difference only in the marker-positive group. Samples are collected from all the subjects to help testing for additional markers in the future.

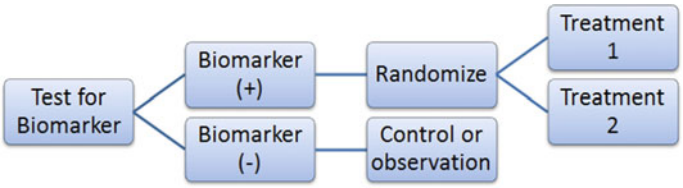


Fig. 2.4 Hybrid design

It should be noted that except for the enrichment/targeted design, all classical designs covered are all-comers designs so that all patients irrespective of their biomarker status is enrolled in the study.

2.3.2 Novel Designs

2.3.2.1 Adaptive Accrual Design

If biomarker-based subgroups are predefined, but with uncertainty on the best possible endpoint and population, an adaptive accrual design could be considered with interim analysis that may lead to modify the patient population to accrual.

Wang et al. (2007) proposed a phase III design comparing an experimental treatment with a control treatment that begins with accruing both positive and negative biomarker status patients. An interim futility analysis would be performed, and based on results of the interim analysis it is decided to either continue the study in all patients or only the biomarker positive patients. Specifically, the trial follows the following scheme: begin with accrual to both marker-defined subgroups; an interim analysis is performed to evaluate the test treatment in the biomarker-negative patients. If the interim analysis indicates that confirming the effectiveness of the test treatment for the biomarker-negative patients is futile, then the accrual of biomarker-negative patients is halted and the final analysis is restricted to evaluating the test treatment for the biomarker-positive patients. Otherwise, accrual of biomarker-negative and biomarker-positive patients continues to the target sample size until the end of the trial. At that time, the test treatment is compared to the standard treatment for the overall population and for biomarker-positive patients (Fig. 2.5).

Jenkins et al. (2011) proposed a similar design but with more flexibility in the context of oncology trials. It allows the trial to test treatment effect in the overall population, subgroup population or the co-primary populations at the final analysis based on the results from interim analysis. Besides, the decision to extend to the second stage is based on intermediate or surrogate endpoint correlated to the final endpoint (Fig. 2.6). Specifically, the trial has two distinct stages and follows the following scheme: at the first stage, accrual both marker-defined subgroups; an interim analysis is performed on the first stage subjects using a short-term intermediate endpoint;

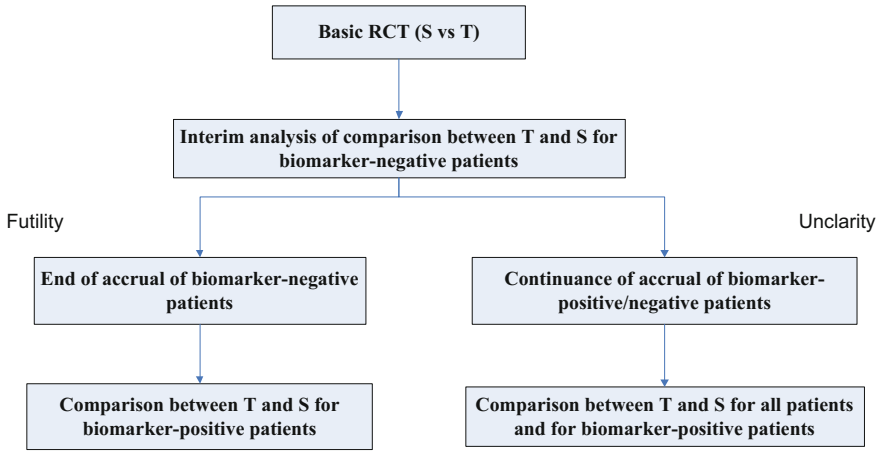


Fig. 2.5 Adaptive accrual design

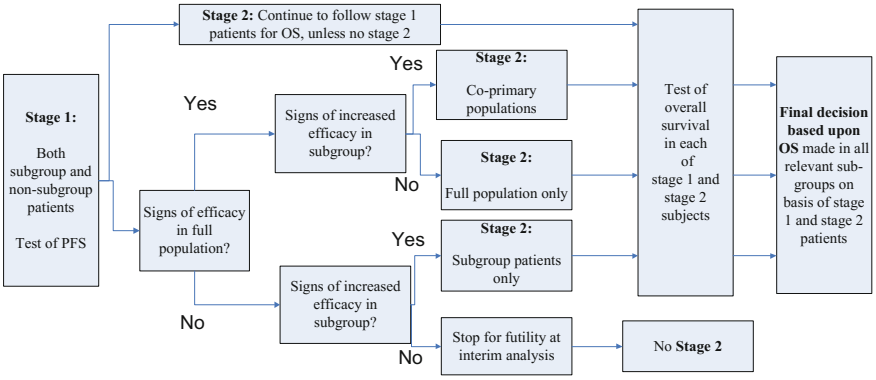


Fig. 2.6 Adaptive accrual design

based on the interim results, the trial can (1) continue in co-primary populations; (2) continue in marker-defined subgroup; (3) continue in the full population without an analysis in marker-defined subgroup; (4) stop for futility. Each of the above options has a pre-specified, but potentially different, stage 2 sample size and length of follow up associated with it. As the trial continues to recruit new subjects for stage 2, the stage 1 subjects would be remained in the trial and be monitored for long-term endpoint. The final assessment for the trial is on long-term endpoint for all patients from both stages.

Mehta et al. (2009) also proposed an adaptive accrual design, which is called "adaptive group sequential design with population enrichment." This design is an extension to a classical group sequential design. Specifically, at the second-last visit in a trial using group sequential design, both the sample size and study population are allowed to be modified based on the accumulated observed data. Adaptive

accrual designs are very attractive due to its flexibility to change sample size and enrich population, which greatly increase the chance of study success. However, these designs also increase the complexity of trial dramatically. From the trial management perspective, logistics of drug are complicated if sample size is increased and the recruitment might slow down if population is enriched. From the statistical perspective, type I error of the trial would be inflated due to the potential interim adaptations and multiplicity. Appropriate statistical correction methods and testing procedures should be applied to preserve the type I error. Besides, intensive simulations should be conducted to obtain a good understanding of the various trial parameters such as the interim decision rule before committing to a design.

2.3.2.2 Biomarker-Adaptive Threshold Design

Biomarker development and validation is usually very expensive and time consuming. Often times by the time of the start of late phase clinical trials, a reliable biomarker, as well as its threshold, for identifying patients sensitive to an experimental treatment is not known.

When the marker is known but the threshold or the cut-point for defining a positive or negative biomarker status is not clear, a biomarker-adaptive threshold design can be considered (Jiang et al. 2007). The biomarker-adaptive threshold design combines the test of overall treatment effect with the establishment and validation of a cut-point for a pre-specified biomarker which identifies a biomarker-based subgroup believed to be most sensitive to the experimental treatment. This design potentially provides substantial gain in efficiency.

Specifically, the main purpose of the biomarker-adaptive threshold design is to identify and validate a cutoff point for a pre-specified biomarker, and to compare the clinical outcome between experimental and control treatments for all patients and for the patients identified as biomarker positive in a single study. The procedure provides a prospective statistical test of the hypotheses that the experimental treatment is beneficial for the entire patient population or that the experimental treatment is beneficial for a subgroup defined by the biomarker and provides an estimate of the optimal biomarker cutoff point.

The statistical hypothesis test can be carried out by splitting the overall type I error rate α. First, compare the treatment response on the overall population at $\alpha 1$ and if not significant then perform the second test at $\alpha - \alpha 1$. For example, if the null hypothesis of no benefit in overall population is rejected at a desired significance level of say 0.04, then the testing is stopped. Otherwise, the testing is carried out at 0.01 to test the hypothesis of no benefit in identified biomarker-based subpopulation. This strategy controls overall alpha below the 0.05 level. The advantage of this procedure is its simplicity and that it explicitly separates the effect of the test treatment in the broad population from the subgroup specification. However, it takes a conservative approach in adjusting for multiplicity in combining the overall and subgroup analyses. Other strategies of combining the two statistical tests for overall and subgroup patients involve consideration of the correlation structure of

the two test statistics. A point estimate and a confidence interval for the cutoff value could be estimated by a bootstrap re-sampling approach.

2.3.2.3 Adaptive Signature Design

The adaptive signature design (Freidlin et al. 2009) is a design proposed to select the subgroup using a large number of potential biomarkers. This design is appropriate when both the potential biomarkers and the cutoff are unknown; however, there is some evidence that the targeted therapy may work in some of the shortlisted biomarkers.

It combines a definitive test for treatment effect in entire patient population with identification and validation of a biomarker signature for the subgroup sensitive patient population. There are three elements in this design: (a) trial powered to detect the overall treatment effect at the end of the trial; (b) identification of the subgroup of patients who are likely to benefit to the targeted therapy at the first stage of the trial; (c) statistical hypothesis test to detect the treatment difference in sensitive patient population based only the subgroup of patients randomized in the latter half of the trial. These elements are pre-specified prospectively.

Statistical tests should be conducted appropriately in this design to account for multiplicity. A proposed strategy is as follows: test the initial null hypothesis of no treatment benefit in overall population at a slightly lower significance level than the overall alpha of 0.05 (for example, 0.04). If the initial null hypothesis is rejected at the lower significance level, then the targeted therapy is declared superior than the control treatment for the overall population. The hypothesis testing and analysis is complete at this stage. If the first hypothesis is not rejected, then the signature component of the design is used to select a potentially promising biomarker subgroup. It is done by the following steps: split the study population into a training sub-sample and a validation sub-sample of patients. Training sub-sample is used to develop a model to predict the treatment difference between targeted therapy and control as a function of baseline covariates. The developed model is then applied to validation sub-sample to obtain the prediction for each subject in this sample. A predicted score is calculated to classify the subject as sensitive or non-sensitive. The subgroup is selected using a pre-specified cutoff for this predicted score. The second hypothesis test is conducted in this sensitive subgroup to see the benefit of the targeted therapy against the control. This test is conducted at a much lower significance (e.g., 0.01).

According to Freidlin and Simon (2009), this design may be ideal to use for Phase II clinical trials for developing signatures to identify patients who respond better to targeted therapies. The advantage of this design is its ability to de-risk losing the label of broader population. However, since only half of the patients are used for development or validation, and with the large number of potential biomarkers for consideration, a large sample size may be needed to adequately power the trial.

2.3.2.4 Cross-Validated Adaptive Signature Design

Cross-validated adaptive signature design (Freidlin et al. 2010) is an extension of the adaptive signature design, which allows use of entire study population for signature development and validation.

Similar to the adaptive signature design, the initial null hypothesis is to test the benefit of the targeted therapy against the control is conducted in the overall population which is conducted at a slightly lower significance level $\alpha 1$ than the overall alpha α. The sensitive subset is determined by developing the classifier using the full population. It is done by the following steps

1. Test the initial null hypothesis of no treatment benefit in the overall population at $\alpha 1$, which is a slightly lower significance level than the overall α. If this hypothesis is rejected, then the targeted therapy is declared superior than the control treatment for the overall population and analysis is completed. If the first hypothesis is not rejected; then carrying out the following steps for signature development and validation.
2. Split study population into "k" sub-samples.
3. One of the "k" sub-samples is omitted to form a training sub-sample. Similar to the adaptive signature design, develop a model to predict the treatment difference between targeted therapy and control as a function of baseline covariates using training sub-sample. Apply the developed model to each subject not in this training sub-sample so as to classify patients as sensitive or non-sensitive.
4. Repeat the same process leaving out a different sample from the "k" sub-samples to form training sub-sample. After "k" iterations, every patient in the trial will be classified as sensitive or non-sensitive.
5. Compare the treatment difference within the subgroup of patients classified as sensitive using a test statistic (T). Generate the null distribution of T by permuting the two treatments and repeating the entire "k" iterations of the cross-validation process. Perform the test at α-$\alpha 1$. If the test is rejected, then the superiority is claimed for the targeted therapy in the sensitive subgroup.

The cross-validation approach can considerably enhance the performance of the adaptive signature design as it permits the maximization of information contributing to the development of the signature, particularly useful in the high-dimensional data setting where the sample size is limited. Cross-validation also maximizes the size of the sensitive patient subset used to validate the signature. One drawback is the fact that the signature for classifying sensitive patients in each sub-sample might not be the same and thus can cause difficulty in interpreting the results if a significant treatment effect is identified in the sensitive subgroup.

2.3.2.5 Basket and Umbrella Trial Designs

Increasing knowledge about the genetic causes of disease is prompting intense interest in the concept of precision medicine. This is particularly the case in oncology,

which researchers view as the field most advanced with the strategy. The science is prompting researchers to develop treatments that target the mutations regardless of where a patient's cancer is located in the body.

A key driver of the strategy is the fact that the same cancer-causing molecular traits are often found in a variety of tumor types, raising hope that a drug effective against the target in, say breast cancer, would be effective in a tumor originating in another organ. Indeed, Roche Holding AG's breast cancer drug Herceptin, which targets a receptor called Her2, turned out to be effective—and was eventually approved—for gastric tumors that have high levels of Her2. But the drug Zelboraf, which is especially effective against the skin cancer melanoma with a certain mutation in a gene called BRAF, turns out to have essentially no effect against colon cancer harboring the same mutation, raising the issue that it is much more complicated and researchers should have some caution toward broad success in the approach.

Another major issue in the clinical development of precision medicines is that genetically characterized tumors breaks common cancers such as lung or breast into a dozen or more much rarer diseases. That poses a challenge to drug companies, which in recruiting for a single-drug trial could have to screen as many as 10,000 patients to find enough patients to test a drug against a rare mutation. Screening patients for a trial involving 10 or 20 drugs instead is expected to be much more efficient, and more quickly provide patients with access to potentially beneficial treatments.

Umbrella trial design and basket trial design are proposed in recent years to meet these challenges and to develop novel targeted therapies in a faster and more efficient manner.

As shown in Fig. 2.7, an umbrella trial assesses different molecularly targeted drugs on different mutations in one cancer type of histology. Examples are Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging And molecular Analysis 2 (I-SPY TRIAL 2, I-SPY 2, NCT01042379; Ref. Barker et al. 2009), the FOCUS4 study in advanced colorectal cancer (Kaplan et al. 2007), and the phase II adaptive randomization design Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination (BATTLE; Ref. Kim et al. 2011) in NSCLC (NCT00409968).

A basket trial assesses one or more molecularly targeted drugs on one or more mutations regardless of cancer types of histologies. This design facilitates a particular targeted therapeutic strategy (i.e., inhibition of an oncogenically mutated kinase) across multiple cancer types. Examples are NCI's Molecular Analysis for Therapy Choice (MATCH) and the Molecular Profiling-based Assignment of Cancer Therapeutics (MPACT, NCT01827384) trials (Conley et al. 2014).

These designs are quite powerful because they can screen and test multiple treatments, multiple biomarkers in multiple indications simultaneously.

The NCI-MATCH basket trial is illustrated below as an example of basket design.

NCI-MATCH: Molecular Analysis for Therapy Choice, announced at the 2015 annual meeting of the American Society of Clinical Oncology (ASCO) in Chicago, is a large basket trial initiated by the National Cancer Institute (NCI). The trial seeks to determine whether targeted therapies for people whose tumors have specific

Fig. 2.7 Umbrella trial design and basket trial design

gene mutations will be effective regardless of their cancer type. NCI-MATCH will incorporate more than 20 different study drugs or drug combinations, each targeting a specific gene mutation, in order to match each patient in the trial with a therapy that targets a molecular abnormality in their tumor.

NCI-MATCH is a phase II trial with numerous small substudies (arms) for each treatment being investigated. It opened with approximately 10 substudies, moving to 20 or more within months. The study parameters for the first 10 arms would be sent to 2,400 participating sites in the NCTN for review in preparation for patient enrollment beginning in July 2015. Additional substudies could be added over time as the trial progresses.

The NCI-MATCH trial has two enrollment steps. Each patient would initially enroll for screening in which samples of their tumor will be removed (biopsied). The samples would undergo DNA sequencing to detect genetic abnormalities that may be driving tumor growth and might be targeted by one of a wide range of drugs being studied. If a molecular abnormality is detected for which there is a specific substudy available, to be accepted in NCI-MATCH patients would be further evaluated to determine if they meet the specific eligibility requirements within that arm. Once enrolled, patients would be treated with the targeted drug regimen for as long as their tumor shrinks or remains stable. Potential treatments include (but not limited to) the following:

- Crizotinib—Separate studies in ALK rearrangements and ROS-1 translocations
- Dabrafenib and trametinib—BRAF V600E or V600K mutations
- Trametinib—BRAF fusions or non-V600E, non-V600K BRAF mutations
- Afatinib—Separate studies in EGFR and HER2 activating mutations

- AZD9291—EGFR T790M and rare EGFR activating mutations
- T-DM1—HER2 amplifications
- VS-6063—NF2 loss
- Sunitinib—cKIT mutations.

Overall, 3000 patients will be screened during the full course of the NCI-MATCH trial to enroll about 1000 patients in the various treatment arms. Each arm of the trial will enroll up to 35 patients. For every trial, the primary endpoint is objective response. Secondary endpoints include 6-month progression-free survival, time to progression, toxicity, and biomarker status.

2.3.3 Examples

2.3.3.1 Example 1: Development of Crizotinib in ALK+ NSCLC

Lung cancer is currently the leading cause of cancer death in both men and women. Historically, lung cancer was categorized as two types of diseases: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), with NSCLC accounting for about 85–90% of lung cancer cases. NSCLC can also be classified according to histological type: adenocarcinoma, squamous-cell carcinoma, and large-cell carcinoma. Such classification is important for determining management and predicting outcomes of the disease.

However, with the rapid advance in biological and genetic science in the past two decades, researchers find there are various oncogenic drivers behind the progression of lung cancer caused by the inactivation of the so-called tumor suppressor genes. Table 2.1 shows the potential oncogenic drivers in NSCLC based on the current knowledge, such as the EGFR mutation and the ALK mutation.

The MTA crizotinib (XALKORI®, Pfizer Inc., New York, NY, USA) is a potent, selective, small-molecule competitive inhibitor of anaplastic lymphoma kinase (ALK), MET, and ROS-1 (Christensen et al. 2007; Shaw et al. 2014). The first-in-human phase 1 trial started in December 2015 to estimate the MTD opening to all-comer patients with solid tumors. The EML4-ALK translocation in NSCLC was discovered in 2007. In the same year, the study was amended to add patients with EML4-ALK mutation to the MTD cohort, and the first clinical response was observed in ALK+ tumors in early 2008. Subsequently, the clinical development program progressed rapidly, and crizotinib was approved in 2011 by the FDA for NSCLC that is ALK-positive as detected by an FDA-approved companion diagnostic test, a commercially available break-apart fluorescence in situ hybridization (FISH) probes for detecting ALK gene rearrangement to detect the rearrangement in NSCLC (Kwak et al. 2010). It took only 6 years from FIH to registration.

Table 2.2 summarizes the clinical studies and their trial designs and endpoints that led to the accelerated and full approval by the global health authorities. Classical enrichment designs were used for these studies that allowed for the investigation of

Table 2.1 Oncogenic drivers in lung adenocarcinoma

Oncogenic drivers	Prevalence (%)
KRAS	20–25
EGFR	13–17
ALK	3–7
MET skipping	~3
HER2	~2
BRAF	~2
PIK3CA	~2
ROS1	~1
MET amp	~1
NRAS	~1
MEK	~1
AKT	~1
RET	~1
NTRK1	~0.5

Table 2.2 Clinical studies that led to accelerated approval and full approval (www.clinicaltrials.gov)

Protocol	Setting	Trial design	Primary endpoints
A8081001 (n = 119)	All lines, solid tumors, ALK-positive NSCLC	Single-arm, open-label study of crizotinib	Safety, pharmacokinetics, response
A8081005 (n = 136)	≥2nd line ALK-positive NSCLC	Single-arm, open-label study of crizotinib	Safety, response
A8081007 (confirmatory phase 3) (n = 318)	2nd line ALK-positive NSCLC	Crizotinib versus (pemetrexed or docetaxel), randomized, open-label study	PFS

this novel drug in an efficient and rapid way because patients with ALK mutation only account for approximately 5% of NSCLC population. High response rates (55–60%) in the two single-arm enrichment studies led to accelerated approval by the FDA. Full approval was granted after positive readout of randomized confirmatory study A8081007.

In the absence of comparative data, it was unclear whether the distinct clinicopathologic characteristics of patients with ALK-positive NSCLC noted above might be contributing to the observed antitumor activity of Crizotinib. Extensive retrospective statistical analyses were conducted using bootstrapping (covariate-matched) and modeling (covariate-adjusted) to simulate outcomes of randomized controlled studies of crizotinib versus standard advanced NSCLC treatment (Selaru et al. 2016).

These analyses utilized data from the control arms of three Pfizer-sponsored phase III studies evaluating first-line paclitaxel—carboplatin or gemcitabine—cisplatin and second- or later-line erlotinib regimens in patients with advanced unselected NSCLC. These analyses demonstrated clinically meaningful and statistically significant effect of Crizotinib despite the lack of a concurrent active control arm.

2.3.3.2 Example 2: Bayesian Predictive Probability Design for an Enrichment Phase 2 POC Study

Breast cancer is a common type of cancer among women. A diagnosis of triple-negative breast cancer (TNBC) means the three most common types of receptors that fuel cancer growth—ER, PR, and HER2 are not present, which represents 15% of breast cancer patients. In TNBC with Notch genomic alternations (NA+), the inhibition of activation of the Notch pathway using single-agent Notch inhibitor therapy may induce clinical activity (Stylianou et al. 2006). The prevalence of Notch alteration in breast cancer is estimated to be around 10%, so Notch+ TNBC represents only 1–2% of breast cancer, a very rare disease.

This is a phase 2 proof-of-concept (POC) study of an experimental Notch inhibitor, an oral drug given twice daily (BID). The hypothesis is that treatment with this drug response rate can be improved from historical level of $\leq$30 to $\geq$60%. However, there are two main challenges in designing the trial. First of all, there is no prior clinical data to suggest a high response rate of 60% can be achieved in this rare disease defined by NA+, nor is there any prior data on the analytical validity or clinical utility of the assay. As a result, it is highly desirable to stop the trial early if observed objective response rate (ORR) is low during the trial conduct. Secondly, due to the extremely low prevalence rate of 1–2% of the target population in breast cancer, enrollment speed is expected to be slow, albeit 20–25 sites will be opened to screen hundreds of TNBC patients, and the turnaround time of the next generation sequencing assay (~2–3 weeks) may decrease trial acceptance.

To meet the aforementioned challenges, a Bayesian predictive probability (BPP) design was proposed with multiple interim looks (Lee and Liu 2008), so that the trial can be stopped early if there is no or low drug effect since it is a costly study with high risk. The Bayesian approach allows greater flexibility in continuously monitoring the trial data to make a go/no-go decision.

Figure 2.8 illustrates the study design. Patients would be tested for the biomarker status. If it is NA positive, patients will be assigned to the experimental drug using the proposed BPP design. It is estimated that at least 28 patients are required to test the hypothesis controlling for the type I and type II error rates. Also 20 NA negative patients would be enrolled to gather some data for exploratory analysis to meet regulatory requirement of health authorities (but not hypothesis to be tested). This is because the treatment effect is expected to be much smaller (if there were any effect) in the marker-negative population, the size of the marker-negative population would usually be too small to give a definitive answer on the effect in that population; but, it would provide at least some estimate of the effect in that population.

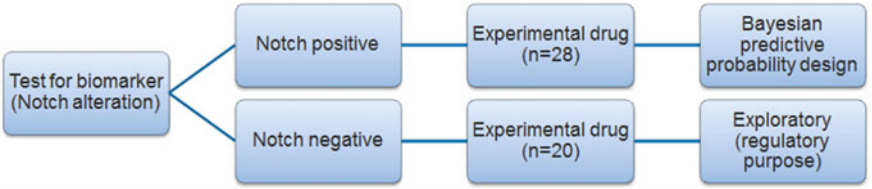


Fig. 2.8 Study design of an experimental notch inhibitor

Table 2.3 Interim futility/efficacy decision rules

Number of patients	Negative/Futility	Positive/Efficacy
8	≤1	≥8
10	≤2	≥9
12	≤3	≥10
15	≤5	≥11
18	≤6	≥12

The BPP design uses a beta-binomial conjugate distribution for the tumor response rate p. The predictive probability (PP) is the probability of a positive result at the end of the trial, based on the cumulative information in the current stage (statistically, an integration of conditional power over the parameter space of p). During the trial, the PP is compared to some boundary values (θ_L and θ_U) for futility and superiority evaluation.

- If pp$<\theta_L$, stop the trial and reject the alternative
- If pp$>\theta_U$, stop the trial and reject the null; otherwise continue.

By applying some optimization algorithms, the optimal design that minimizes the maximum sample size can be determined.

With a non-informative prior of beta (0.3, 0.7), it is estimated that 28 patients will be required to have 25 response-evaluable patients so as to control the one-sided type I error rate at 0.05 with 90% power when the true ORR is 60%. The design has multiple interim looks for potential early stopping, and the decision rules are provided in Table 2.3. At the final analysis, at least 12 responders are required out of 25 evaluable patients to claim the drug efficacious.

It is assumed that futility boundaries are binding. The nominal type I error rate (one-sided) is 0.041 assuming the futility boundaries are binding. Simulations (details not included) show that the pre-specified type I error rate (0.05) is strictly controlled under the non-binding assumption and bounded above at 0.044.

The efficacy boundaries are non-binding which means if crossed enrollment will continue, and the study will not be stopped early. However, this information may form the basis for internal decision making and strategic planning (e.g., opening additional sites to shorten the time to study completion, or starting early discussions with health authorities on potential registration). Non-binding efficacy boundaries will not inflate the type I error rate, although it reduces the power slightly.

With the proposed Bayesian design, the study has a 78% probability of early termination under the null hypothesis, and the expected sample size under the null hypothesis is 14.43 patients. This design achieves the objective of terminating the trial early with minimal resources when the effect of the drug is not as high as expected.

2.3.3.3 Example 3: Adaptive Group Sequential Design with Population Enrichment for Cardiovascular Research

Despite substantial progress in the prevention of cardiovascular disease and its ischemic complications, it remains the single largest killer in the USA. New treatment options are needed, particularly to respond to the challenges of an aging population and rising rates of obesity and diabetes. Development of novel therapeutic strategies for the management of acute cardiovascular disease is especially challenging. Specific problems include relatively low event rates, diverse patient populations, lack of reliable surrogate end points, and small treatment effects subject to substantial uncertainty. Because the clinical development process is enormously expensive and time consuming, there is considerable interest in statistical methods that use accumulating data from a clinical trial to inform and modify its design. Such redesign might include changes in target sample size and even changes in the target population (Mehta 2009). Mehta et al. (2009) demonstrated how to apply an adaptive group sequential design with population enrichment in cardiovascular research.

Consider a placebo-controlled randomized cardiovascular trial with a composite primary end point including death, myocardial infarction, or ischemia-driven revascularization during the first 48 h after randomization for therapies intended to reduce the risk of acute ischemic complications in patients undergoing percutaneous coronary intervention. Assume, based on prior knowledge that the placebo event rate is in the range of 7–10%. The investigational drug is assumed, if effective, to reduce the event rate by 20%, but the evidence to support this assumption is limited. The actual risk reduction could be larger but could also easily be as low as 15%, a treatment effect that would still be of clinical interest given the severity and importance of the outcomes, but that would require a substantial increase in sample size.

Assume, besides the entire population under study G0, two subgroups of patients, G1 and G2, have also been identified by investigators that of interest. G1 is a subset of G0, and G2 is a subset of G1. And it is preferred that the sample size of the study would not exceed 15,000.

A classical group sequential design was considered first with target sample size 8750 patients and interim analyses be performed at 50 and 70% of that target (Fig. 2.9). The first interim look would take place when the first 4375 patients have finished the study, and the trial stops for efficacy if the test statistic $Z1 \leq -2.96$ and for futility if $Z1 \geq 0.1$. The second interim look would take place when 6215 patients have finished the study, and the corresponding efficacy and futility bounds for Z2 are -2.46 and -1.29, respectively. At the final look, when all 8750 patients have completed the study, the null hypothesis of no treatment effect is rejected in favor

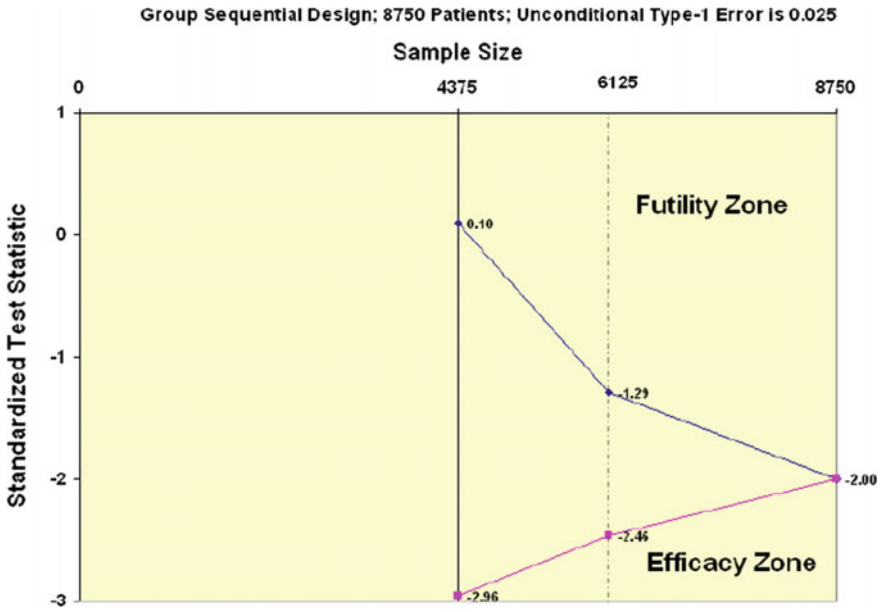


Fig. 2.9 Classical group sequential design

of the active treatment if $Z3 \leq -2.0$. Otherwise, we fail to reject the null hypothesis. The unconditional type 1 error is controlled at 0.025 for this group sequential design.

Assume at the first and second looks, we observed the test statistics, $Z1 = -1.6728$ and $Z2 = -1.8816$ (Fig. 2.10), both do not exceed either the efficacy or futility boundaries. Therefore, the trial can continue beyond the second look. Besides, based on the observed data up to the second interim look, the estimated event rate for the control arm is 8.7% and the estimated percent risk reduction is only 15.07%. For these values, the conditional power of this study is only 67%.

To increase the conditional power to 80%, as it is always preferred, the sample size needs to be increased. Besides, as the total sample size is preferred not to exceed 15,000, the design could be adapted as follows:

(1) If the sample size reestimation for the overall population G0 produces a revised sample size >15,000, consider enrichment.
(2) The enrichment strategy proceeds as follows: (i) estimate the number of additional patients needed to test the null hypothesis of no treatment effect in G1 with 80% conditional power, assuming that the observed effect size in G1 is the true effect size; (ii) if that sample size plus the number of patients already enrolled is <15,000, the trial will continue until the additional number of patients is enrolled, but future eligibility will be restricted to patients belonging to subgroup G1; (iii) if that enrichment strategy does not yield a sample size <15,000, the same calculation will be performed with the estimated effect size for subgroup G2 and, if that reestimation yields a sample size <15,000, the trial will

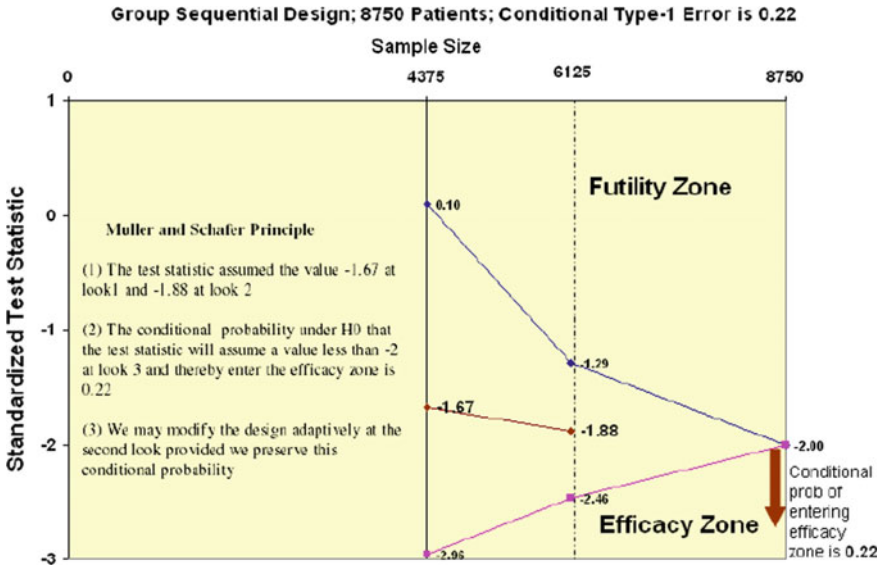


Fig. 2.10 Group sequential design with observed values

continue, enrolling only members of G2; (iv) if neither of the sample size calculations for the patient subgroups yields a sample size <15,000, the trial will be continued with the original eligibility criteria and sample size target, provided that the conditional power with 8750 patients is at least 20%; (v) otherwise, the trial will be terminated for futility.

A question arises as to how to test for a treatment effect at the end of the trial given the possibility of a sample size increase and population enrichment at the second interim look. To preserve the type 1 error of the study as a whole, a closed testing procedure that guarantees strong control of the type 1 error rate can be employed. Specifically, if enrichment is implemented, the testing procedure involves 2 hypothesis tests (Fig. 2.11). Suppose for specificity that the data lead to an enrichment strategy with the G1 subpopulation. In that instance, test 1 is a test of the null hypothesis in the entire study population, consisting of patients enrolled from G0 before the initiation of the enrichment strategy and patients enrolled from G1 thereafter. However, if test 1 rejects the null hypothesis, we can conclude only that the new treatment is superior to placebo in either the G0 population or the G1 subpopulation. Thus, we then perform test 2, a conventional level-α test of the null hypothesis of no treatment effect in patients from the G1 subpopulation enrolled after the second interim analysis, that is, after enrichment began. If both hypotheses are rejected, we conclude that treatment is superior to placebo in the subpopulation G1. In this way, the family-wise error rate is strongly controlled. A similar procedure is followed if enrichment is limited to subpopulation G2.

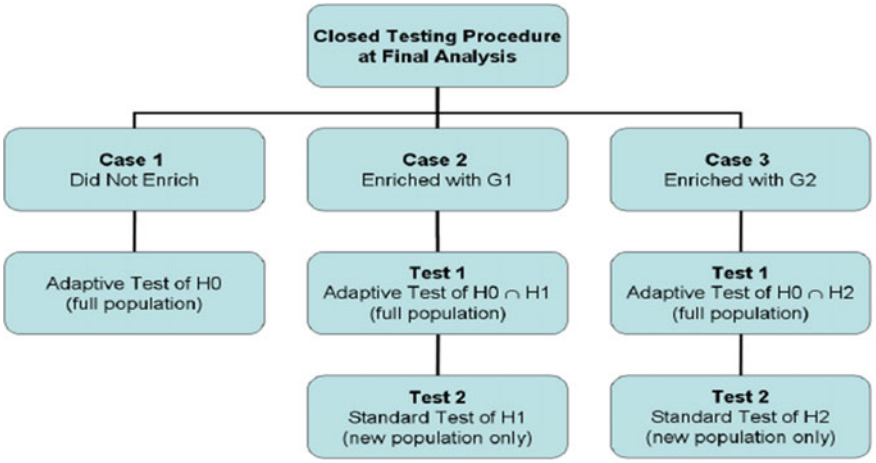


Fig. 2.11 Closed testing procedure *Note* Hi: no treatment effect in patient population Gi

2.4 Discussion

Omics biomarkers have become increasingly important in understanding disease biology and providing guidance to clinical trial designs. Such information has been widely incorporated in clinical trials, especially in oncology trials due to the heterogeneity nature of the disease, which is often defined by many types of omics biomarker. Leveraging the biomarker information, numerous innovative designs have been proposed in the literature and have presented tremendous opportunities for innovation in drug development. Benefits of advanced biomarker-informed designs include, but not limited to, (i) higher probability of success with more accurate targeting population defined by biomarker; (ii) the flexibility to modify the trial to gain clinical benefits; (iii) the possibility to shorten the development cycle; (iv) ability to leverage more data outside of the trial. However, steep challenges such as predictability of biomarker and robust biomarker assay development remain key issues to be addressed before the implementation of advanced designs involving biomarker. The novel biomarker-informed designs should be viewed as "a" solution not "the" solution for planning clinical trial experiment, and the use of the novel design should be fully evaluated and applied depending on the context. Thorough statistical simulations are encouraged to be in place before any decision making. In principle, any novel design can only be implemented "by design" and the statistical validity and integrity must be preserved. Guidelines regarding principles of how and when to use novel designs should be developed such that the risk of misuse and misinterpretation of the novel designs are kept at minimum.

References

Albert, S. R., Sinicrope, F. A., & Grothey, A. (2005). N0147: A randomized phase III trial of oxaliplatin plus 5-fluorouracil/leucovorin with or without cetuximab after curative resection of stage III colon cancer. *Clinical Colorectal Cancer, 5*(3), 211–213.

Amado, R. G., Wolf, M., Peeters, M., Van Cutsem, E., Siena, S., Freeman, D., et al. (2008). Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *Journal of Clinical Oncology, 26*(10), 1626–1634.

Baker, S. G., & Freedman, L. S. (1995). Potential impact of genetic testing on cancer prevention trials, using breast cancer as an example. *Journal of the National Cancer Institute, 87,* 1137–1144.

Baker, S. G., & Kramer, B. S. (2005). Statistics for weighing benefits and harms in a proposed genetic substudy of a randomized cancer prevention trial. *Applied Statistics, 54*(5), 941–954.

Baker, S. G., Kramer, B. S., Sargent, D. J., & Bonetti, M. (2012). Biomarkers, subgroup evaluation, and clinical trial design. *Discovery medicine, 13*(70), 187–192.

Barker, A. D., Sigman, C. C., Kelloff, G. J., Hylton, N. M., Berry, D. A., & Esserman, L. J. (2009). ISPY2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics, 86,* 97–100.

Bauer, P. (1991). Multiple testing in clinical trials. *Statistics in Medicine, 10,* 871–890.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289–300.

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., et al. (2016). Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinformatics, 17*(Suppl 2), S15.

Bi, R., & Liu, P. (2016). Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics, 17,* 146.

Bokemeyer, C., Bondarenko, I., Makhson, A., Hartmann, J. T., Aparicio, J., de Braud, F., et al. (2009). Fluorouracil, leucovorin, and oxaliplatin with and without cetuximab in the first-line treatment of metastatic colorectal cancer. *Journal of Clinical Oncology, 27*(5), 663–671.

Bonetti, M., & Gelber, R. D. (2000). A graphical method to assess treatment-covariate interactions using the cox model on subsets of the data. *Statistics in Medicine, 19,* 2595–2609.

Bonetti, M., & Gelber, R. D. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics, 5,* 465–481.

Cai, T., Tian, L., Wong, P. H., & Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics, 12*(2), 270–282.

Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., et al. (2011). Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *New England Journal of Medicine, 364*(26), 2507–2516.

Christensen, J. G., et al. (2007). Cytoreductive antitumor activity of PF-2341066, a novel inhibitor of anaplastic lymphoma kinase and c-Met, in experimental models of anaplastic large-cell lymphoma. *Molecular Cancer Therapeutics, 6,* 3314–3322.

Cobo, M., Isla, D., Massuti, B., et al. (2007). Customizing cisplatin based on quantitative excision repair cross-complementing 1 mRNA expression: A phase III trial in non-small-cell lung cancer. *Journal of Clinical Oncology, 25*(19), 2747–2754.

Conley, B. A., & Doroshow, J. H. (2014). Molecular analysis for therapy choice: NCI MATCH. *Seminars in Oncology, 41,* 297–299.

Fadista, J., Manning, A., Florez, J., & Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics, 24,* 1202–1205.

Freidlin, B., McShane, L. M., & Korn, E. L. (2010). Randomized clinical trials with biomarkers: Design issues. *Journal of the National Cancer Institute, 102*(3), 152–160.

Freidlin, B., Jiang, W., & Simon, R. (2009). The cross-validated adaptive signature design. *Clinical Cancer Research, 16*(2), 691–698.

Hong, E. P., Park, J. W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics & Informatics, 10*(2), 117–122.

Janes, H., Pepe, M. S., Bossuyt, P. M., & Barlow, W. E. (2011). Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine, 154*(4), 253–259.

Jenkins, M., Stone, A., & Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoint. *Pharmaceutical Statistics, 10*(4), 347–356.

Jiang, W., & Yu, W. (2016). Power estimation and sample size determination for replication studies of genome-wide association studies. *BMC Genomics, 17*(Suppl 1), 3.

Jiang, W., Freidlin, B., & Simon, R. (2007). Biomarker-adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute, 99*(13), 1036–1043.

Johnstone, I., & Titterington, D. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A, 367,* 4237–4253.

Jung, S., & Young, S. (2012). Power and sample size calculation for microarray studies. *Journal of Biopharmaceutical Statistics, 22*(1), 30–42.

Kaplan, R., Maughan, T., Crook, A., Fisher, D., Wilson, R., Brown, L., et al. (2007). Evaluating many treatments and biomarkers in oncology: A new design. *Journal of Clinical Oncology, 31,* 4562–4568.

Karapetis, C. S., Khambata-Ford, S., Jonker, D. J., O'Callaghan, C. J., Tu, D., Tebbutt, N. C., et al. (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine, 359*(17), 1757–1765.

Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein, G. R., Jr., Tsao, A., et al. (2011). The BATTLE trial: Personalizing therapy for lung cancer. *Cancer Discovery, 1,* 44–53.

King, M. C., Wieand, S., Hale, K., Lee, M., Walsh, T., Owens, K., et al.; National Surgical Adjuvant Breast and Bowel Project. (2001). Tamoxifen and breast cancer incidence among women with inherited mutations in BRCA1 and BRCA2: National surgical adjuvant breast and bowel project (NSABP-P1) breast cancer prevention trial. *JAMA, 286*(18), 2251–2256.

Klein, R. J. (2007). Power analysis for genome-wide association studies. *BMC Genetics, 8,* 58.

Kwak, E. L., et al. (2010). Anaplastic lymphoma kinase inhibition in non–small-cell lung cancer. *New England Journal of Medicine, 363,* 1693–1703.

Lee, J., & Liu, D. (2008). A predictive probability design for phase II cancer clinical trials. *Clinical Trials, 5*(2), 93–106.

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics, 13,* 762–775.

Liu, P., & Hwang, J. T. G. (2007). Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics, 23*(6), 739–746.

Lowe, W. L., & Reddy, T. E. (2015). Genomic approaches for understanding the genetics of complex disease. *Genome Research, 25,* 1432–1441.

Mandrekar, S. J., & Sargent, D. (2009). Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. *Journal of Clinical Oncology, 27*(24), 4027–4034.

Mehta, C., Gao, P., Bhatt, D. L., Harrington, R. A., Skerjanec, S., & Ware, J. H. (2009). Optimizing trial design. *Circulation, 119*, 597–605.

Meienberg, J., Bruggmann, R., Oexle, K., Matyas, G. (2016). Clinical sequencing: Is WGS the better WES? *Human Genetics, 135*, 359–362.

Panagiotou, O. A., & Ioannidis, J. P. (2012). Genome-wide significance project. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology, 41*(1), 273–286.

Park, J. H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., et al. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics, 42,* 570–575.

Pe'er, I., Yelensky, R., Altshuler, D., Daly, M. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology, 32*, 381–385.

Pineda, S., Real, F. X., Kogevinas, M., Carrato, A., Chanock, S. J., Malats, N., et al. (2015). Integration analysis of three omics data using penalized regression methods: An application to bladder cancer. PLoS Genetics, *11*(12).

Romond, E. H., Perez, E. A., Bryant, J., Suman, V. J., Geyer, C. E. Jr, Davidson, N. E., et al. (2005). Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *New England Journal of Medicine, 353*(16), 1673–1684.

Saccenti, E., & Timmerman, M. E. (2016). Approaches to sample size determination for multivariate data: Applications to PCA and PLS-DA of omics data. *Journal of Proteome Research, 15,* 2379–2393.

Sargent, D. J., Conley, B. A., Allegra, C., & Collete, L. (2005). Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology, 23*(9), 2020–2227.

Selaru, P., Tang, Y., Huang, B., Polli, A., Wilner, K., Donnelly, E., et al. (2016). Sufficiency of single-arm studies to support registration of targeted agents in molecularly selected patients with cancer: Lessons from the clinical development of Crizotinib. *Clinical and Translational Science*.

Shabalin, A., Tjelmeland, H., Fan, C., Perou, C., & Nobel, A. (2008). Merging two gene-expression studies via cross-platform normalization. *Bioinformatics, 24*(9), 1154–1160.

Shaw, A. T., et al. (2014). Crizotinib in ROS1-rearranged non-small-cell lung cancer. *New England Journal of Medicine, 371,* 1963–1971.

Shyr, D., & Liu, Q. (2013). Next generation sequencing in cancer research and clinical application. *Biological Procedures Online, 15,* 4.

Simon, R., & Wang, S. J. (2006). Use of genomic signatures in therapeutics development in oncology and other diseases. *The Pharmacogenomics Journal, 6*(3), 166–173.

Song, X., & Pepe, M. S. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics, 60*(4), 874–883.

Song, Y., & Chi, G. Y. (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine, 26*(19), 3535–3549.

Spencer, C. C., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics, 5,* e1000477.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, 64,* 479–498.

Stylianou, S., Clarke, R., & Brennan, K. (2006). Aberrant activation of notch signaling in human breast cancer. *Cancer Research, 66*(3), 1517–1525.

Thompson, J., Tan, J., & Greene, C. (2016). Cross-platform normalization of microarray and RNA-seq data for machine learning applications. *Peer J, 4,* e1621.

Van Cutsem, E., Köhne, C. H., Hitre, E., Zaluski, J., Chang Chien, C. R., Makhson, A., et al. (2009). Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *New England Journal of Medicine, 360*(14), 1408–1417.

Wang S. J., O'Neill R. T., & Hung, H. M. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics, 6*(3), 227–244.

Wetterstrand, K. A. (2016). DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP). Retrieved 23 December 2016, from www.genome.gov/sequencingcostsdata.

Wu, Z., & Zhao, H. (2009). Statistical power of model selection strategies for genome-wide association studies. *PLoS Genetics, 5,* e1000582.

Chapter 3
Phase 3 Oncology Trials of Personalized Medicines with Adaptive Subpopulation Selection

Cong Chen, Wen Li, Xiaoyun (Nicole) Li and Robert A. Beckman

3.1 Introduction

A personalized medicine in oncology may benefit a subpopulation of patients with certain predictive biomarker signatures or certain disease types. However, limited by preclinical data and early phase clinical data, there is great uncertainty about drug activity in a subpopulation when designing a confirmatory trial in practice and it is logical to take a two-stage adaptive approach. The first stage de-selects (or prunes) non-performing subpopulations at an interim analysis, and the second stage pools the remaining subpopulations in the final analysis. There are two important designs in this context. In the first design (Li et al. 2016, 2017), patients with different biomarker levels are enrolled in a study and the treatment effect is assumed to be in ascending order of the biomarker level (a biomarker enrichment design). The goal of the interim analysis is to de-select non-performing biomarker subpopulations. In the second design (Chen et al. 2016a; Yuan et al. 2016; Beckman et al. 2016), patients with different tumor types but the same biomarker signature are included in a trial (a basket design). The goal of the interim analysis is to identify a subset of tumor types in the absence of order in treatment effect.

Two-stage designs represent a special type of adaptive designs. Most of the previous research work in adaptive designs has used the same endpoint for both interim and final analyses (Stallard and Todd 2010; Magnusson and Turnbull 2013; Mehta et al. 2014). However, in practice it often takes a long time to observe the primary

C. Chen (✉) · W. Li · X. (Nicole) Li
Biostatistics and Research Decision Sciences, Merck & Co, Inc., Kenilworth, NJ, USA
e-mail: cong_chen@merck.com

R. A. Beckman
Departments of Oncology and of Biostatistics, Bioinformatics, and Biomathematics, Lombardi Comprehensive Cancer Center and Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington, DC 20007, USA
e-mail: eniac1915@gmail.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_3

endpoint of a study (e.g., overall survival in oncology), and the data on a primary endpoint at an interim analysis may provide little information on the true treatment effect. To gather more informative data in a typical confirmatory oncology trial, it often requires majority of (if not all) patients to have an adequate follow-up, which makes it difficult to implement a timely adaptation. One solution to this problem would be to de-select subpopulations at the end of the trial, using the primary endpoint and a subset of the patients, aka, the "informational design" (Chen et al. 2016b). When an intermediate endpoint sensitive to treatment intervention is available, an alternative solution is to use it at the interim analysis for an adaptation decision (See Chen et al. (2013) for a general discussion of the utility of a sensitive intermediate endpoint in late-stage drug development). In oncology, such an endpoint is routinely derived from tumor size changes post-treatment, either as a continuous or a categorical variable.

When an intermediate endpoint is used at interim to assist with the design, a key issue is how to incorporate it into Type I error control and estimation of treatment effect for the primary endpoint. Stallard (2010) used a double regression method to incorporate intermediate endpoint data into the estimation of treatment effect on the primary endpoint, which improved the accuracy of population selection. Type I error control was demonstrated via simulations. Royston et al. (2003) explored the issue in non-confirmatory trials without specifically controlling Type I error. Jenkins et al. (2011) proposed a design that uses a combination test for population selection and hypothesis testing based on two different endpoints. Their proposed design involves separating enrolled subjects into two distinct partitions (a typical approach in adaptive designs). No explicit assumptions were made about the treatment effect on intermediate endpoint in these approaches. Wang et al. (2014) and Friede et al. (2011) proposed different adaptive designs for population selection and hypothesis testing. In their approaches, the intermediate endpoint at the interim analysis is assumed to have no treatment effect in Type I error control. Wang et al. (2013) investigated the impact of the correlation between the two endpoints on Type I error control. Again, the intermediate endpoint at the interim analysis is assumed to have no treatment effect.

Since confirmatory Phase 3 trials routinely fail on primary endpoints in spite of clinical activities observed on intermediate endpoints in Phase 1 or Phase 2 trials, it is inappropriate to assume null treatment effect on an intermediate endpoint when discussing Type I error control of the primary endpoint. The possible existence of a mild to moderate effect should be considered. In this chapter, we treat the treatment effect on an intermediate endpoint as a nuisance parameter to provide the most conservative Type I error control and explore the maximum bias of the naïve estimator of the treatment effect. The handling of intermediate treatment effect as a nuisance parameter is a distinctive feature of our proposed methodology. Note that the use of same endpoint can be viewed as a special case within the framework of this chapter.

The remaining chapter is organized as follows. Section 3.2 introduces notations. Sections 3.3 and 3.4 review the control of overall Type I error and treatment effect estimation for the biomarker enrichment design and the basket design, respectively. A hypothetical example is provided under each design. In the hypothetical examples, we

have kept the total sample size of the pooled population the same as planned without pruning (refer to Chen et al. (2016a) for other sample size adjustment strategies), which requires a sample size increase of the performing subpopulations as necessary. However, the methodology presented in this chapter does not rely on any particular sample size adjustment strategy. Section 3.5 provides summary and discussions.

3.2 Statistical Designs of a Phase 3 Trial with Biomarker or Tumor Selection

We consider a population composed of K distinct subpopulations. To demonstrate whether an experimental therapy (compared to control) has a treatment effect in a subset of the subpopulations, a two-stage trial is planned under a Phase 3 setting. Since time-to-event endpoints are the most popular in Phase 3 oncology trials, for illustration purpose, progression-free survival (PFS) is used as an intermediate endpoint for population selection at the interim analysis while overall survival (OS) is the primary endpoint for the final analysis. When one or both endpoints are not of the time-to-event type, the results presented in this chapter can be easily extended once a joint distribution function of the involved test statistics is obtained.

For simplicity, interim analysis is assumed to be conducted at a common information time t for all the subpopulations in this chapter. The information time refers to the proportion of target number of deaths for final analysis. A non-performing subpopulation is de-selected or pruned after the interim analysis if it does not meet a pre-specified selection criterion. Assume that m subpopulations are identified as non-performing and the remaining (K-m) subpopulations are deemed to be performing based on the selection criteria. (Of note, m is not pre-specified.) The (K-m) performing ones proceed to the second stage and are pooled as one composite population in the final analysis. If all the subpopulations are identified as non-performing, the study stops early for futility. For simplicity, a common selection bar (i.e., p value for testing treatment effect in a subpopulation $< \alpha_t$) is applied to all the subpopulations. A trend of the observed treatment effect favoring the experimental arm may correspond to $\alpha_t < 0.5$, and a potentially clinically significant finding may correspond to $\alpha_t < 0.025$.

A key issue of interest for a study design is Type I error control in hypothesis testing of the primary endpoint at the final analysis. When patients used in the interim analysis are excluded from the final analysis (an operational adaption), there is not any concern on Type I error inflation. However, this is not a preferred approach in practice because the final analysis would be substantially delayed and the sample size of the study might be substantially increased. Throughout this chapter, we consider an inferential adaptive approach in that patients from the first stage are included in the final analysis. A penalty on Type I error control has to be paid as a consequence. We use α^* to denote the nominal alpha level at the final analysis that maintains the overall Type I error at an α level (usually 0.025 one-sided) acceptable to regulatory

agencies. Another key issue of interest is the potential bias of the naïve estimator of treatment effect based on the pooled population at the final analysis (Bauer et al. 2010). Before we discuss the two issues, we introduce the notations and basic distributional properties of the relevant test statistics.

3.2.1 Notations and Distributional Properties

The true treatment effect of the time-to-event endpoints refers to negative logarithm of hazard ratio (experimental arm versus control arm) and is denoted by δ_i for the intermediate endpoint at interim and by θ_i for the primary endpoint in the end for the *i*th subpopulation, $1 \le i \le K$. The asymptotic distributions of the test statistics based on the two endpoints are as follows:

- The standardized test statistics $X_i(t)$ based on the intermediate endpoint used at the interim analysis for population selection follows a normal distribution $N\left(\delta_i\sqrt{I_i(t)}, 1\right)$, where $I_i(t)$ is the Fisher information of the intermediate endpoint at the interim time t (i.e., number of events on the intermediate endpoint divided by 4);
- The standardized test statistics $\tilde{X}_i$ based on the primary endpoint used at the final analysis follows a normal distribution $N\left(\theta_i\sqrt{\tilde{I}_i}, 1\right)$, where $\tilde{I}_i$ is the Fisher information of the primary endpoint at the final analysis (i.e., number of events on the primary endpoint divided by 4).

Furthermore, we have $\mathrm{Corr}(X_i(t), X_{i'}(t)) = 0$, $\mathrm{Corr}\left(\tilde{X}_i, \tilde{X}_{i'}\right) = 0$ for $i \neq i'$. To simplify the discussion, we assume that the two test statistics for each subpopulation have a common correlation $\rho(\ge 0)$, i.e., $\mathrm{Corr}\left(X_i(t), \tilde{X}_i\right) = \rho$, $1 \le i \le K$. The higher the correlation, the higher the potential inflation Type I error. Overall Type I error rate is asymptotically controlled as long as a consistent estimator of ρ is used for the calculations introduced below. With this simplification, the standardized test statistics has a multivariate normal distribution with a simple correlation structure that is not impacted by the information although in general it may. The correlation can be estimated using the trial data by the WLW method (Wei et al. 1989). Alternative methods such as the bootstrapping method (Hall et al. 1989) may also be used to estimate the correlation structure. When the endpoints are of different types, the standard test statistics have the canonical joint normal distribution as defined in Jennison and Turnbull (2000), which can be estimated accordingly.

3.2.2 Pooled Treatment Effect in the Performing Subpopulations

At the end of the trial, the (K-m) performing subpopulations are pooled together for the primary analysis through a standardized test statistics $V_{(-m)}$ which can be written as

$$V_{(-m)} = \sum_{j=m+1}^{K} \sqrt{\tilde{I}_j}\tilde{X}_j \Bigg/ \sqrt{\sum_{j=m+1}^{K} \tilde{I}_j} \tag{3.1}$$

The test statistics $V_{(-m)}$ follows an asymptotic normal distribution, i.e., $V_{(-m)} \sim N\left(\sum_{j=m+1}^{K} \theta_j \tilde{I}_j \Big/ \sqrt{\sum_{j=m+1}^{K} \tilde{I}_j}, 1\right)$. When the Fisher information of the primary endpoint at the final analysis ($\tilde{I}_i$) is equal for all the performing subpopulations, $V_{(-m)}$ can be rewritten as $\sum_{j=m+1}^{K} \tilde{X}_j / \sqrt{K-m}$.

Let S_m denote a set of (K-m) performing subpopulations that continue after the interim analysis, a point estimate of the corresponding pooled treatment effect can be written as

$$\hat{\theta}_{S_m} = \frac{V_{(-m)}}{\sqrt{\sum_{j=m+1}^{K} \tilde{I}_j}} \tag{3.2}$$

In the following sections, S is used to denote a generic set of the performing subpopulations, and $\hat{\theta}_S$ is used to denote the corresponding treatment effect.

3.3 Population Section in a Biomarker Enrichment Design

Without loss of generality, we assume that the treatment effect from the first subpopulation to the *K*th subpopulation is in an ascending order. At the interim analysis, a subpopulation i is identified as performing if its standardized test statistic $X_i(t)$ is no less than $Z_{1-\alpha_t}$ and is identified as non-performing otherwise, where $Z_{1-\alpha_t}$ is the lower $(1-\alpha_t)$th quantile of a standard normal distribution. The examination of non-performing subpopulations begins from the first subpopulation and continues up until the first performing subpopulation (denoted as the $(m+1)$th subpopulation) is identified ($m \geq 0$). The subpopulations starting from $(m+1)$th are judged as "performing" and are pooled in the final analysis.

3.3.1 Control of Type I Error

According to the design descriptions above, the overall Type I error is

$$\begin{aligned} & P\left(\text{Reject the null hypothesis at the level } \alpha^* | \{\theta_i = 0 \delta_i : i = 1, \ldots, K\}\right) \\ & = P\left(X_1(t) > Z_{1-\alpha_t}, V_{(-0)} > Z_{1-\alpha^*} | \{\theta_i = 0, \delta_i : i = 1, \ldots, K\}\right) \\ & + \sum_{m=1}^{K-1} P\left(X_1(t) \le Z_{1-\alpha_t}, \ldots, X_m(t) \le Z_{1-\alpha_t}, X_{m+1}(t) > Z_{1-\alpha_t}, \right. \\ & \qquad \left. V_{(-m)} > Z_{1-\alpha^*} | \{\theta_i = 0, \delta_i : i = 1, \ldots, K\}\right) \end{aligned} \tag{3.3}$$

Equation (3.3) can be calculated from a multivariate normal distribution based on the distributional results in Sect. 3.2. In order to control the overall Type I error, the probability (i.e., the overall Type I error) in (3.3) must be no more than the pre-specified α level.

The calculation of (3.3) involves the nuisance parameters δ_i ($1 \le i \le K$). Since the trial data are not necessarily generated under $\theta_i = 0$ ($1 \le i \le K$), it is impossible to estimate δ_i ($1 \le i \le K$) from the trial data. However, in most disease settings, it is reasonable to assume that δ_i tends to be closer to zero when θ_i is under the null hypothesis than under an alternative hypothesis. This assumption can be helpful for narrowing the parameter space for δ_i ($1 \le i \le K$). In addition, a meta-analysis of relevant historical trials may also be used to find a sensible parameter space. Both approaches are out of the scope of this chapter. Instead, we try to find the minimal α^* that controls (3.3) at α in the entire parameter space $(-\infty, \infty)$. The minimal α^* (denoted as $\min(\alpha^*)$) is used as the nominal alpha level in the final analysis. As expected, $\min(\alpha^*)$ is less than α in general (see technical details in Li et al. 2016).

3.3.2 Treatment Effect Estimation

Implied by the fact that $\min(\alpha^*)$ is less than α, the naïve estimator $\hat{\theta}_S$ may overestimate the treatment effect. Given that $X_{m+1}(t)$ has a left truncated normal distribution conditional on S_m, the bias of the point estimate $\hat{\theta}_{S_m}$ at the final analysis, denoted as Bias_{m+1}, can be shown to be

$$\frac{\sqrt{\tilde{I}_{m+1}}\rho\phi\left(Z_{1-\alpha_t} - \delta_{m+1}\sqrt{I_{m+1}(t)}\right)/\left(1 - \Phi\left(Z_{1-\alpha_t} - \delta_{m+1}\sqrt{I_{m+1}(t)}\right)\right)}{\sum_{j=m+1}^{K} \tilde{I}_j}.$$

The overall bias of the point estimate $\hat{\theta}_S$ is

$$\text{Bias}_{\text{overall}} = \sum_{m=0}^{K-1} \text{Bias}_{m+1} \times P(S = S_m) \tag{3.4}$$

where $P(S = S_m)$ is the probability of de-selecting m non-performing subpopulations, which is $\left(\prod_{j=1}^{m} \Phi\left(Z_{1-\alpha_t} - \delta_j\sqrt{I_j(t)}\right)\right) \times \left(1 - \Phi\left(Z_{1-\alpha_t} - \delta_{m+1}\sqrt{I_{m+1}(t)}\right)\right)$ for $m>0$ and $\left(1 - \Phi\left(Z_{1-\alpha_t} - \delta_1\sqrt{I_1(t)}\right)\right)$ for $m=0$.

Similar to Type I error control, a conservative estimate of the bias can be obtained by maximizing (3.4) in the entire parameter space of δ_i for $1 \le i \le K$. Unlike in Type I error control whereas δ_i cannot be estimated for the value that corresponds to the null hypothesis $\theta_i = 0$ $(1 \le i \le K)$, estimation of (3.4) is not tied to an untenable hypothesis, and point estimates of δ_i $(1 \le i \le K)$ based on trial data can be plugged into the equation to derive a less conservative point estimate of the bias. The accuracy of the analytic form (3.4) was demonstrated in a simulation study in Li et al. (2017).

3.3.3 A Hypothetical Example

In this section, we present a hypothetical trial example. Assume a total of 500 patients are randomized with 1:1 ratio into the experimental arm and the control arm. The overall population consists of two biomarker subpopulations (biomarker positive (i.e., BM+) and biomarker negative (i.e., BM−)). OS is the primary endpoint, and PFS is the intermediate endpoint in this study. Enrollment period is 1 year, and accrual rate is constant. Median OS and median PFS in the control arm are assumed to be 1 year and 6 months, respectively. The hazard ratio (HR) between the experimental arm and the control arm in the overall population is 0.75 for OS and is 0.6 for PFS for planning purpose. The study would complete after a total of 360 deaths are observed, which is approximately 1.5 years after enrollment completion. The study has about 78% power to detect 0.75 hazard ratio in OS at $\alpha = 0.025$.

To implement our proposed adaptive design, the interim analysis is planned when all patients are enrolled. At the time of the interim analysis, approximately 260 PFS events are observed in the overall population. For simplicity, we assume that there are equal numbers of events in each biomarker subpopulation for both OS and PFS. We choose $\alpha_t = 0.1$ at the interim analysis for population selection; i.e., the observed HR in PFS would be less than approximately 0.8 in order for a biomarker subpopulation to be included in the final analysis. The sample size in BM+ will have to be increased in case BM− is de-selected at interim in order to reach the target number deaths.

We assume that the correlation between the log-rank test statistics based on PFS at interim and the log-rank test statistics based on OS at final is estimated to be 0.5, a reasonable if not over-conservative estimate based on our experience. With this correlation estimate, $\min(\alpha^*)$ is calculated to be 0.0187 based on the study setup and (3.3) and is obtained at $\delta_1 = 0.2$ and any $\delta_2 \ge 1.7$. Consider a scenario when true HR for OS is 0.56 in BM+ and 1.00 in BM−, and true HR for PFS is 0.45 in BM+ and 1.00 in BM−. In this scenario, our proposed design can successfully de-select the BM− subpopulation 90% of the time at the interim analysis and have an overall study power of 95%. We also evaluated the power of using OS for de-selection at the same interim analysis under $\alpha_t = 0.2$. In this case, $\min(\alpha^*)$ is calculated to be 0.0219

based on our setup and a degenerate form of (3.3). The corresponding study power is 90%, lower than the design that uses PFS for population selection. Overall, our proposed design, using either the intermediate endpoint or the clinical endpoint, has a higher power than the traditional design without subpopulation selection (78%).

Based on (3.4), the maximum bias in the naïve estimator of the treatment effect is 0.009. With the true hazard ratio on OS at around 0.75, this bias leads to an estimated hazard around 0.74. The maximum bias is obtained at $\delta_1 = 0.125$ and $\delta_2 = 0.225$.

3.4 Tumor Type Section in a Basket Design

The basket design de-selects a subpopulation (i.e., a tumor type) to the second stage whenever its standardized test statistic $X_i(t)$ is less than the selection bar $Z_{1-\alpha_t}$. Similar to the biomarker enrichment design described in the previous section, the remaining subpopulations are pooled in the final analysis.

3.4.1 Control of Type I Error

Suppose m subpopulations are pruned at the interim analysis, and they are denoted by $i_1, \ldots, i_m$. The remaining (K-m) performing subpopulations are denoted by $i_{(m+1)}, \ldots, i_K$. In this situation, $V_{(-m)}$ is similarly defined as in (3.1), but is based on $X_{i_{(m+1)}}(t), \ldots, X_{i_K}(t)$. With m non-performing subpopulations pruned at interim, the overall Type I error is $P_0\left(\alpha^*, i_1, \ldots, i_m | \alpha_t, m, \left\{\theta_{i_j} = 0, \delta_{i_j} : j = 1, \ldots, K\right\}\right)$ or equivalently $P\left(X_{i_1}(t) \le Z_{1-\alpha_t}, \ldots, X_{i_m}(t) \le Z_{1-\alpha_t}, X_{i_{(m+1)}}(t) > Z_{1-\alpha_t}, \ldots, X_{i_K}(t) > Z_{1-\alpha_t}, V_{(-m)} > Z_{1-\alpha^*} | \left\{\theta_{i_j} = 0, \delta_{i_j} : j = 1, \ldots, K\right\}\right)$.

In the absence of the order of treatment effect by subpopulation, there are C(K, m) possible configurations of $\{i_1, \ldots, i_m\}$ where $C(K, m) = C(K, m) = K!/((K-m)!m!)$. The overall Type I error for basket design using an intermediate endpoint is in the following form

$$P\left(\text{Reject the null hypothesis at the level } \alpha^* | \{\theta_i = 0, \delta_i : i = 1, \ldots, K\}, t\right) = \sum_{m=0}^{K-1} \sum_{\text{over } C(K,m)^*} P_0\left(\alpha^*, i_1, \ldots, i_m | \alpha_t, m, \left\{\theta_{i_j} = 0, \delta_{i_j} : j = 1, \ldots, K\right\}\right) \quad (3.5)$$

where the second summation is over all the C(K, m) possible configurations. Equation (3.5) can be calculated from a multivariate normal distribution based on the distributional results in Sect. 3.2.

Similar to the biomarker enrichment design, a narrow parameter space for δ_i $(1 \le i \le K)$ based on trial data or a meta-analysis may be used to facilitate the calculation of α^*. To be conservative, we search the entire parameter space of δ_i $(1 \le i \le K)$ to find the minimal α^* that controls (3.5) at α. Due to the symmetry of the test statis-

tics under the null hypothesis as well as lack of ordering in tumor type selection, the minimal α^* is achieved at same $\vartheta_i = \delta_i\sqrt{I_i(t)}$ for $1 \le i \le K$, which can simplify (3.5) substantially. The minimal α^* (denoted as $\min(\alpha^*)$) is used for testing the primary endpoint in the final analysis. As expected, $\min(\alpha^*)$ is less than α in general, which again implies possible bias in point estimation of treatment effect.

3.4.2 Treatment Effect Estimation

Based on a similar derivation in Sect. 3.3, the overall bias is

$$\sum_{m=0}^{K-1} \sum_{\text{over } C(K,m)*} \left(\text{Bias}_{m+1} \times P\left(S = S_m = \left\{i_{(m+1)}, \ldots, i_K\right\}\right)\right) \tag{3.6}$$

where $\text{Bias}_{m+1} = \frac{\sum_{j=m+1}^{K} \sqrt{\tilde{I}_{i_j}}\,\rho\left(\phi\left(Z_{1-\alpha_t} - \delta_{i_j}\sqrt{I_{i_j}(t)}\right) \Big/ \left(1 - \Phi\left(Z_{1-\alpha_t} - \delta_{i_j}\sqrt{I_{i_j}(t)}\right)\right)\right)}{\sum_{j=m+1}^{K} \tilde{I}_{i_j}}$,

and $P\left(S = S_m = \left\{i_{(m+1)}, \ldots, i_K\right\}\right)$ is $\left(\prod_{j=1}^{m} \Phi\left(Z_{1-\alpha_t} - \delta_{i_j}\sqrt{I_{i_j}(t)}\right)\right) \times \prod_{j=m+1}^{K}\left(1 - \Phi\left(Z_{1-\alpha_t} - \delta_{i_j}\sqrt{I_{i_j}(t)}\right)\right)$ for $m > 0$ and $\prod_{j=1}^{K}\left(1 - \Phi\left(Z_{1-\alpha_t} - \delta_{i_j}\sqrt{I_{i_j}(t)}\right)\right)$ for $m = 0$.

Similar to the biomarker enrichment design, an estimate of the maximum bias can be obtained and so is a less conservative point estimate of the bias based on trial data.

3.4.3 A Hypothetical Example of Adaptive Trial with Tumor Type Selection

In this hypothetical example, the key objective of the basket trial is to file for accelerated approval at an interim analysis (following existing regulatory paradigm) and full approval at the final analysis. PFS is often an acceptable endpoint for accelerated approval. Different endpoints (e.g., ORR for immunotherapies) may also be considered in practice upon discussion with regulatory agencies. A more general objective of an interim analysis is to discontinue the non-performing tumor types earlier.

Consider a randomized controlled basket trial with 1:1 randomization in six tumor indications. Each tumor indication has approximately 88 PFS events at the interim analysis. With 88 events, each indication has 90% power for detecting a hazard ratio of 0.5 in PFS at $\alpha_t = 0.025$ (1-sided). An observed hazard ratio of 0.66 approximately meets the selection bar of $\alpha_t = 0.025$ for a subpopulation to be included in the pooled analysis.

It is assumed that approximately 110 patients are enrolled for each tumor indication in order to reach the target number of PFS event at the data cutoff date for the

interim analysis. The final sample size in the pooled population is fixed at approximately 660 (i.e., 110 multiplied by 6). The interim analysis for each tumor indication is conducted separately unless the target number of PFS events is reached at approximately the same time. Whenever an indication has a negative PFS outcome (i.e., the indication is pruned), a total of 110 patients are equally distributed to the remaining indications. Based on the above setup and (3.5), the minimal α^* is 0.008 when ρ is 0.5. A larger penalty is paid in this example than in the last one, mainly because of the lack of ordering in treatment effect in subpopulations in this example. Assuming that 65% of the patients in the pooled population die by the cutoff date for the final analysis (i.e., a total of 430 events), the study has approximately 90% power to detect a hazard ratio of 0.7 in OS at $\alpha^* = 0.008$. An observed hazard ratio of 0.79 approximately meets the nominal alpha level $\alpha^* = 0.008$ for a positive OS outcome in the pooled analysis.

As a comparison to a traditional design without pruning at the interim analysis, the observed hazard ratio in OS would be approximately 0.83 for the trial to be positive at $\alpha = 0.025$. While this bar seems easier to cross, when some of the tumor indications in the basket are inactive, the chance of crossing it in a trial that does not prune non-performing ones at interim is lower than an adaptive trial that does (Chen et al. 2016a). Moreover, the addition of an interim analysis for accelerated approval could substantially shorten the time to drug approval. Last but not least, this basket trial has the potential to get an experimental therapy approved in up to six tumor indications based on a single trial with comparable sample size to a conventional Phase 3 trial for one tumor indication.

When $\rho = 0.5$, the maximum bias is estimated to be approximately 0.07. This means that, when the true hazard ratio is 0.75, the observed hazard ratio is expected to be approximately 0.70 (i.e., $\exp(-(-\log(0.75) + 0.07))$) in the worst-case scenario for the underlying intermediate treatment effect. The actual bias can be estimated based on estimates of the PFS effects at the interim analysis. Of note, the maximum bias is the same under different selection bars. However, for a certain range of δ (say a more reasonable range of 0.4–0.6), the corresponding bias could be different under different selection bar, e.g., a smaller bias yielded by a relaxed bar $\alpha_t = 0.1$. This is because the greater the α_t, the less the cherry picking effect in pruning and the less the estimation bias.

3.5 Discussion

We have provided a statistical approach to control the overall Type I error and correct estimation bias when different endpoints are used for adaption and hypothesis testing in a two-stage adaptive design setting. The approach is illustrated with two previously studied designs useful for the development of personalized medicines. The same approach can be applied to more general adaptive designs. Related issues are the focus of ongoing research.

The Type I error control in this chapter is based on a null hypothesis of all subpopulations being inactive. Rejection of the null hypothesis does not automatically mean that the subpopulations in the pool are equally active. Heterogeneity in treatment effect across subpopulations in the final analysis deserves investigation. However, this is a common issue in conventional Phase 3 trials. For example, the impact of baseline characteristics on treatment effect is routine investigated in Phase 3 trials. Regulatory decision on drug approval or scope of the label hinges upon the outcome of such ad hoc analyses despite an overall positive outcome from the trial. The issue is also similar to regional effect in a multi-regional study or the trial effect in a meta-analysis, both well studied and understood. The concern about heterogeneity should not be a hurdle of conducting biomarker enrichment trials or basket trials.

For simplicity, we have only considered one interim analysis in the hypothetical examples. In practice, multiple interim analyses may be conducted. For example, an interim analysis for detecting an overwhelming survival benefit may be added after pruning. Timing of the interim analysis for pruning is important for reducing selection errors. The interim analysis may be conducted earlier if treatment effect is expected to manifest early or later otherwise.

We note that the required penalty for using internal data for pruning can be significant especially for the basket trial. An alternative approach is to rely on credible external data for pruning. Further, there should be no penalty when subpopulations are pruned solely because of unexpectedly slow accrual, or due to evolution of standard of care that renders the study obsolete. Relaxation of the selection criteria may also reduce the penalty.

Last but not least, it is well known that the naïve estimator of treatment effect in a conventional group sequential trial that has stopped early for efficacy can be inflated (Wang et al. 2016). Pruning in a two-stage adaptive design has a similar impact on estimation of treatment effect as early stopping in a group sequential design does. However, unlike in the group sequential trial setting whereas there is no data to substantiate the bias estimation, patients enrolled in the second stage of a two-stage adaptive design provide an unbiased data source for this purpose. The data should be routinely reported in practice, along with the bias-corrected estimate.

References

Bauer, P., Koenig, F., Brannath, W., & Posch, M. (2010). Selection and bias—Two hostile brothers. *Statistics in Medicine, 29*(1), 1–13. https://doi.org/10.1002/sim.3716.

Beckman, R. A., Antonijevic, Z., Kalamegham, R., & Chen, C. (2016). Design for a basket trial in multiple tumor types based on a putative predictive biomarker. *Clinical Pharmacology & Therapeutics for publication*. https://doi.org/10.1002/cpt.446.

Chen, C., Sun, L., & Chih, C. (2013). Evaluation of early efficacy endpoints for proof-of-concept trials. *Journal of Biopharmaceutical Statistics, 23,* 413.

Chen, C., Li, N., Yuan, S., Antonijevic, Z., Kalamegham, R., & Beckman, R. A. (2016a). Statistical design and considerations of a Phase 3 basket trial for simultaneous investigation of multiple

tumor types in one study. *Statistics in Biopharmaceutical Research, 8*(3), 248–257. https://doi.org/10.1080/19466315.2016.1193044.

Chen, C., Li, N., Shentu, Y., Pang, L., & Beckman, R. A. (2016b). Adaptive informational design of confirmatory phase III trials with an uncertain biomarker effect to improve the probability of success. *Statistics in Biopharmaceutical Research, 8*(3), 237–247. https://doi.org/10.1080/19466315.2016.1173582.

Friede, T., Parsons, N., Stallard, N., Todd, S., Marquez, E. V., Chataway, J., et al. (2011). Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Statistics in Medicine, 30*(13), 1528–1540. https://doi.org/10.1002/sim.4202.

Hall, P., Martin, M. A., & Schucany, W. R. (1989). Better nonparametric bootstrap confidence intervals for the correlation coefficient. *Journal of Statistical Computation and Simulation, 33*(3), 161–172. https://doi.org/10.1080/00949658908811194.

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Boca Raton, Florida, USA: Chapman & Hall/CRC.

Jenkins, M., Stone, A., & Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics, 10,* 347–356. https://doi.org/10.1002/pst.472.

Li, X., Chen, C., & Li, W. (2016). Adaptive biomarker population selection in phase III confirmatory trials with time-to-event endpoints. *Statistics in Biosciences*. https://doi.org/10.1007/s12561-016-9178-4.

Li, W., Chen, C., Li, X., & Beckman, R. A. (2017). Estimation of treatment effect in two-stage confirmatory oncology trials of personalized medicines. *Statistics in Medicine*. https://doi.org/10.1002/sim.7272

Magnusson, B. P., & Turnbull, B. W. (2013). Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine, 32*(16), 2695–2714. https://doi.org/10.1002/sim.5738.

Mehta, C., Schafer, H., Daniel, H., & Irle, S. (2014). Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Statistics in Medicine, 33*(26), 4515–4531. https://doi.org/10.1002/sim.6272.

Royston, P., Mahesh, K. B. P., & Qian, W. (2003). Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in Medicine, 22*(14), 2239–2256. https://doi.org/10.1002/sim.1430.

Stallard, N. (2010). A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine, 29,* 959–971. https://doi.org/10.1002/sim.3863.

Stallard, N., & Todd, S. (2010). Seamless phase II/III design. *Statistical Methods in Medical Research, 20*(6), 623–634. https://doi.org/10.1177/0962280210379035.

Wang, H., Rosner, G. L., & Goodman, S. N. (2016). Quantifying over-estimation in early-stopped clinical trials and the "freezing Effect" on subsequent research. *Clinical Trials, 13*(6), 621–631.

Wang, J., Chang, M., & Menon, S. (2014). Biomarker-informed adaptive design. In M. Chang (Ed.), *Clinical and statistical considerations in personalized medicine* (pp. 129–148). Boca Raton, FL: CRC Press.

Wang, S. J., Brannath, W., Bruckner, M., Hung, H. M. J., & Koch, A. (2013). Unblinded adaptive statistical information design based on clinical endpoint or biomarker. *Statistics in Biopharmaceutical Research, 5*(4), 293–310. https://doi.org/10.1080/19466315.2013.791639.

Wei, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distribution. *Journal of the American Statistical Association, 84,* 1065–1073. https://doi.org/10.1080/01621459.1989.10478873.

Yuan, S., Chen, A., He, L., Chen, C., Gause, C. K., & Beckman, R. A. (2016). On group sequential enrichment design for basket trials. *Statistics in Biopharmaceutical Research for publication, 8*(3), 293–306. https://doi.org/10.1080/19466315.2016.1200999.

Chapter 4
High-Dimensional Data in Genomics

Dhammika Amaratunga and Javier Cabrera

4.1 Introduction

Technological advances are revolutionizing the pharmaceutical industry. Particularly in early-stage research, there has been a significant paradigm shift. Rather than study a few carefully chosen entities (such as certain proteins or enzymes) as in the past, the thinking now is to explore a very large number of entities (e.g., maybe an entire genome or several thousand proteins) all at once.

This has led to high-dimensional data becoming a common characteristic of early-stage biological research, particularly in genomics, proteomics, and imaging. High-dimensional data are data that are generated when p features are measured on each of n samples, so they can be organized into a $p \times n$ matrix X, with n and p such that p is at least an order of magnitude larger than n (i.e., $n \ll p$). It is this latter attribute that distinguishes this type of data, sometimes referred to as "small n, large p" data or megavariate data, from standard statistical multivariate data, which are similar except that their $n > p$.

As an example, in a study of sialic acid storage diseases, the expression levels of $p = 45101$ genes were measured for $n = 12$ 18-day old mice, of which 6 were wildtype mice and 6 were knockout mice, where "knockout" refers to the fact that these mice had their Slc17A5 gene inactivated (Van Acker et al. 2017; Moechars et al. 2005). The measurements were performed on RNA samples drawn from total brain using Affymetrix Mouse 430 2.0 GeneChips. The objective of the study was to identify differences in gene expression patterns across these two sets of mice, since it is known that there is association between mutations in the Slc17A5 gene, which encodes the

D. Amaratunga (✉)
Princeton Data Analytics LLC, Bridgewater, NJ, USA
e-mail: damaratung@yahoo.com

J. Cabrera
Department of Statistics, Rutgers University, Piscataway, NJ, USA

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_4

protein sialin and sialic acid storage diseases. We will use this data (which we will hereon refer to as the Sialin data) for illustrative purposes in the rest of the paper.

The de facto dimensionality of the data is, of course, p. Conventional data analysis methodologies are either not directly applicable or are unlikely to be effective when dealing with this sort of data since the sample size is considerably less than p. Methods developed for regular machine learning and data mining applications may also encounter difficulties as they could have a tendency to overfit since $n<p$; i.e., they are likely to find spurious patterns in the data. In fact, at first sight, the high dimensionality would seem to present an insurmountable problem, not just for conventional methods, but for any method. However, the essential belief when analyzing a high-dimensional dataset is that it contains patterns of value which reside in much lower (say k) dimensional subspaces, where not only $k<p$ but also $k<n$, in fact, ideally $k=1$ or 2. Of course, neither the low-dimensional representations, nor even k, may be unique. In other words, it is possible that there are several low-dimensional aspects of the data involving different subsets of the p initial features which carry some sort of meaningful information. Finding them (and distinguishing them from spurious patterns) is the major challenge.

In the remainder of the paper, we will outline some methods that have been applied successfully for analyzing high-dimensional data in the genomics arena. The number of techniques that have been proposed is quite substantial, and it is not possible in a short review to outline all of them; therefore, only a few select methods that we and our colleagues have found consistently useful will be presented. Tukey's ideas on exploratory data analysis (Tukey 1977, 1980) play a pivotal role in these methods. Further details of procedures can generally be found either online or in the literature. Amaratunga and Cabrera (2004) and its second edition, Amaratunga et al. (2014), are book-length treatments of this topic, while Amaratunga and Cabrera (2016) is a brief review.

4.2 Visualization

Two-dimensional renderings of the data are useful for initial exploration of the data. One such is the biplot, which is a plot associated with the singular value decomposition. Here X is decomposed as $X=UDV'$, where U is $p \times n$ and column orthogonal, V is $n \times n$ and orthogonal, and D is $n \times n$ and diagonal. If we retain only the two largest values in the diagonal of D (call it D_2) and subset U and V accordingly (call them U_2 and V_2), we can approximate X by $X_2 = A_2 B_2'$ where $A_2 = U_2 D_2^{1/2}$ and $B_2 = V_2 D_2^{1/2}$. The two columns of A_2, when plotted against each other, will offer a two-dimensional rendering of the p features. Analogously, the two columns of B_2, when plotted against each other, will offer a two-dimensional rendering of the n samples. The two plots can be shown in a single diagram called a biplot (Gabriel 1971), so that not only the individual characteristics of the p features and the n samples, but also possible associations between the two can be assessed.

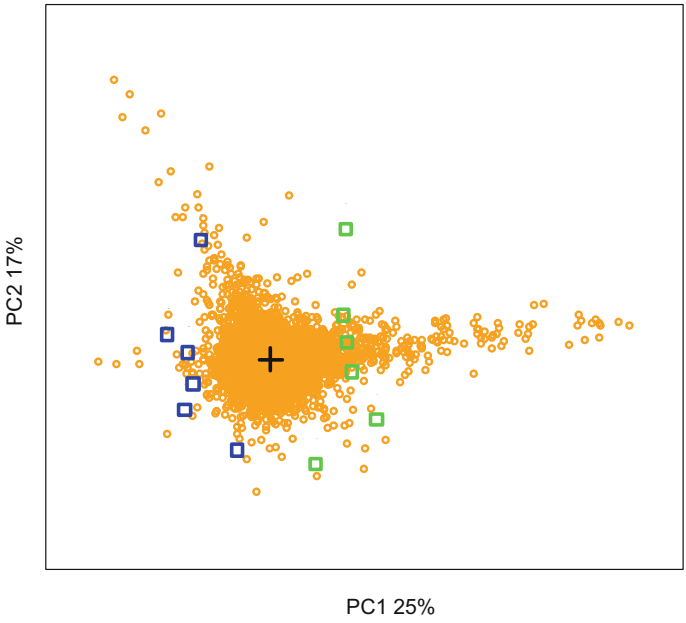


Fig. 4.1 Spectral map of the Sialin data. The 12 samples (the 6 wildtype mice and the 6 knockout mice) are shown as squares, with the wildtype mice shown as blue squares and the knockout mice shown as green squares. The 45101 genes are shown as gold circles

A spectral map is a special type of biplot which is constructed following certain modifications to X, such as adjusting for size and scale differences among the features; for details, see Wouters et al. (2003). This enhances the display for microarray data.

Figure 4.1 is a spectral map of the Sialin data. It can be observed that the two sets of the samples (the wildtype mice and the knockout mice) separate cleanly along the direction of the x-axis, with the wildtype mice (shown as blue squares) to the left and the knockout mice (shown as green squares) to the right. The p value of a t test performed on the x-axis projection of the samples is 3.31×10^{-6}. Especially given that this was an "unsupervised" analysis, in the sense that the construction of the spectral map did not make use of the information that the samples came from two distinct populations, this clearly shows that there is good separation between the samples.

4.3 Individual Feature Analysis

When analyzing grouped high-dimensional data, it is useful to next carry out a supervised analysis of each of the p features individually. This will give an indication as to which features are driving the separation between groups. This can be done using conventional statistical hypothesis testing techniques.

For example, for the Sialin data, since there are two groups of samples, student's t tests can be used. However, the sample size is often limited in these studies, which in turn reduces the power of these tests. Hence, it is often useful to "borrow strength" across features to improve the sensitivity of the entire procedure. This can be done by setting some additional structures, such as setting a distribution structure for the variances.

Let the data be denoted as $\{X_{gij}\}$, where g ($g=1, \ldots, p$) indexes the features, j ($j=1$, 2) indexes the groups, and i ($i=1, \ldots, n_j$) indexes the samples (note: $n_1+n_2=n$). The standard t test assumes that X_{gij} is normally distributed with mean μ_{gj} and variance σ_g^2. Under these assumptions, the t test uses t test statistics $\{T_g\}$ to test whether $\mu_{g1} = \mu_{g2}$ for each feature, g.

A number of suggestions have been made as to how to borrow strength across the p features. Generally, they assume that the $\{\sigma_g^2\}$ collectively follow some distribution F_σ. Most proposed approaches are parametric in nature and assume a distributional form for F_σ, such as an inverse gamma distribution. One widely used such method is *limma* (Smyth 2004). This is a hybrid classical–Bayes approach in which a posterior variance estimate is substituted into the classical t statistic in place of the usual sample variance, giving rise to a moderated t statistic T_g*. Like with the conventional t test, the null distribution of T_g* can be adequately approximated by a t distribution, but with different degrees of freedom.

A semiparametric approach for borrowing strength, which is less dependent on distributional assumptions, is Conditional t or Ct (Amaratunga and Cabrera 2009). In this approach, even the normality assumption of the t test is dropped and it is assumed that X_{gij} follows an unknown distribution F. Now both F and F_σ are unspecified distributions and a resampling scheme along the lines of the bootstrap (Efron 1981) is used to approximate them. They are then used to generate critical envelopes, $t_\alpha(s_g)$ (instead of constant critical values as in the conventional t test) for several different values of α; here s_g is the pooled standard error of the gth feature, and α is the significance level of the test. For each feature, g, a p value, p_g, can be assigned by identifying the smallest value of α that results in significance for that feature.

For the Sialin data, limma declared 990 features as significant at the 0.001 level while Ct declared 909 features as significant at the 0.001 level; these can be considered to be the initial "discoveries" by this analysis.

4.3.1 Multiplicity Considerations

Clearly, however, when testing so many features, the likelihood is very high that there will be a large number of "false discoveries" among the discoveries, i.e., differences that are declared significant even though they are not real. The traditional approach to this multiple testing problems has been to control the probability of at least one false discovery (called the *Familywise Error Rate* or FWER), the Bonferroni procedure being the most common such fix. However, if this is done with a large number of features, there will be a substantial loss of statistical power and many true discoveries

may go undetected. The *False Discovery Rate* (FDR) is a more recent alternative (Benjamini and Hochberg 1995) that seeks to better address this problem.

This approach attempts to control the FDR, defined as:

FDR= expected proportion of false discoveries among the set of discoveries,

rather than the FWER. There are multiple approaches to controlling the FDR, some common ones are by Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), and Storey (2002).

Operationally, FDR control is often done by converting the observed p values to FDR adjusted p values or q values. The FDR adjusted p value (also referred to as a q value) for feature g is the smallest FDR value for which the null hypothesis can be rejected for that feature and all others with smaller p values (Benjamini and Yekutieli 2001; Storey 2002).

For the Sialin data, 338 of the limma features and 345 of the Ct features had q values below 0.001, with an overlap of 305 genes.

4.3.2 Gene Set Analysis

Even with multiplicity adjustments, it is clearly impossible to screen large numbers of apparently significant genes. Thus, it is often of interest to also analyze gene sets, where a gene set, GS, is a collection of genes that are known to share a common biological function, chromosomal location, or regulation. Gene sets are available in public databases such as Gene Ontology (GO), KEGG, Biocarta, and GenMAPP.

The objective of gene set analysis is to identify gene sets which have comparatively low p values (or q values). Thus, it is natural to consider using a version of Fisher's statistic for combining p values (Fisher 1925) for this purpose. This is the Mean Log P (MLP) statistic (Pavlidis et al. 2004; further studied by Raghavan et al. 2006, 2007, Tryputsen et al. 2014):

$$MLP = mean(-\log(p_i)) = \sum_{i \in GS} (-\log(p_i))/\boldsymbol{r}$$

Since gene-level p values often tend to deviate from a uniform distribution even in situations that may be regarded as null, whichever test statistic is used, a permutation procedure is the most effective way to assess whether a given set GS of size r with observed test statistic *MLP* is enriched. This is carried out as follows: repeatedly draw random samples of size r from the set of all p values; recalculate *MLP* for each random sample (call this *MLP**); the proportion of times that *MLP** exceeds the observed value *MLP* in a large number of runs could be regarded as the p value for the significance of gene set GS.

For the Sialin data, one of the more prominent findings is that the GO hierarchy that involves myelination (GO term 42552), ensheathment of neurons (GO term 7272), and regulation of action potential in neurons (GO term 19228) are all significant.

This is highly encouraging from an interpretation point of view since loss of sialin is known to affect these biological processes.

4.4 Analysis of Combinations of Features

Next, it is useful to study combinations of features. Supervised classification (or discriminant analysis) refers to a class of techniques whose objective is to seek a combination of features that is able to discriminate between the groups of samples with reasonable accuracy. Again, there are a number of possible approaches. We shall now describe a method based on fitting a linear model for the case where the number of groups is two.

Let Y_i indicate the group of the ith sample, let π_i be the probability of sample i belonging to Group 1, and let the values of the features for the ith sample be x_i, the ith column of X. The logistic regression model postulates that π_i is associated with x_i via the equation:

$$\log(\pi_i/(1-\pi_i)) = \beta' x_i,$$

where β is a p-vector of coefficients. In conventional logistic regression, these coefficients are estimated by maximizing the log-likelihood:

$$l(\beta) = \sum [Y_i \log(\pi_i) + (1-Y_i)\log(1-\pi_i)]$$

When $n < p$, there is insufficient data to estimate the full model. One solution is to maximize $l(\beta)$ under the penalty constraint $\Sigma\ |\beta_j| < h$, or, equivalently, to minimize

$$S(\beta) = -l(\beta) + \lambda \sum |\beta_j|$$

after scaling all features to have unit sample variance. This procedure is called *lasso* (Tibshirani 1996). The tuning parameter λ controls the strength of the penalty: $\lambda = 0$ yields the standard regression estimates, $\lambda \to \infty$ yields all zero estimates, and values of λ in between these two extremes yield compromises between fitting the traditional logistic model and shrinking all of its coefficients toward zero. A suitable value for λ is usually found by assessing the fit. Once this is done and the model fitted, many coefficients will inevitably shrink all the way to zero, essentially performing feature selection. A highly effective and efficient algorithm for lasso and a related procedure called elastic net was developed by Friedman et al. (2010).

The Sialin dataset was analyzed using lasso. As an initial check to see how good the fit was, the grouping for the 48 samples was predicted using the fitted model; the groups of all 12 were identified correctly; there were zero errors. This may seem to indicate a very good fit, but since the fitting and the prediction were done using

exactly the same data, the zero error rate could be highly over-optimistic. Therefore, it would be unwise to rely on this as an assessment of goodness of fit.

It is best to measure the predictive ability of a model by validating it on a set of data that was not used to fit the model. This could be done by dividing the dataset into two parts, then using one part (called the "training set") for fitting the model and using the other part (called the "test set") for testing it. The group of each sample in the test set can be predicted using the fitted model and the proportion of errors can be calculated to give an assessment of the predictive accuracy of the model. This will not be over-optimistic because the test data were not used for model fitting.

However, there is often not enough data to allow a part of it to be left out for testing. A simulated version of the same idea is *leave-one-out cross-validation* (LOOCV) (Stone 1974). This is carried out in n steps as follows. At the ith step, all the samples except sample i form the training set. The model is fitted using this training set. Then the group of the ith sample, which is now temporarily the test set, is predicted using the fitted model and it is noted whether or not the prediction is correct. This is repeated for each i (i.e., for each sample), in turn and the percentage of total errors is calculated and reported as the LOOCV error rate.

The LOOCV error rate for the Sialin data was 0% (i.e., 0/12), indicating a good fit. We also kept track of which genes appear in at least half the fitted models. These are possibly the genes most influential for separating the two groups, although because of correlations among genes, certain influential ones may not show up. These genes can be examined as before, and again, the GO hierarchy involving GO terms 42552, 7272, and 19228 shows up as an affected pathway.

Variations of cross-validation include leave-k-out cross-validation (in which k samples are left out at each step) and k-fold cross-validation (in which the original set of samples is randomly partitioned into k subsets, one of which is left out at each step). Another type of variation is the bootstrap (in which a random set of n samples is left in at each step, with the random sampling being done with replacement) (Efron 1981, 1983; Breiman 1996). There are variants of the bootstrap too, such as the .632+bootstrap (Efron 1983; Efron and Tibshirani 1997).

When there are a large number of features, ensemble techniques, which iterate through samples of both rows and columns of X with the findings aggregated and collated at the end, have been found to work well. The initial popular ensemble technique was Random Forest (Breiman 2001), which was mostly developed for mining early "Big Data" where n was very large. A variation called Enriched Random Forest (Amaratunga et al. 2008a, b) works well when $n<p$. An additional variation, in which lasso is used in a Random Forest like procedure (Amaratunga et al. 2012) has also been found to work well.

4.5 Discussion

In this paper, we have given an overview of techniques that are useful for analyzing grouped high-dimensional genomics data. When there is no group information, the goal of the analysis might actually be to try and infer groups among the samples.

For this, *unsupervised classification* (also called *cluster analysis*) techniques can be applied. Here too special methods have been developed for high-dimensional data, an example being an ensemble technique called ABC (Amaratunga et al. 2008a, b). Two-way clustering and biclustering methodologies have also been developed and applied (Kasim et al. 2016).

Finally, an important note: it is imperative that any findings be independently validated. Due to the high dimensionality, overfitting always remains a possibility, particularly in the selection of important features. Independent verification maybe sought through a repeat or similar study or through contextual subject-matter means. The importance of such independent qualification cannot be stressed enough.

References

Amaratunga, D., & Cabrera, J. (2004). *Exploration and analysis of DNA microarray and protein array data*. New York: John Wiley.

Amaratunga, D., & Cabrera, J. (2009). A conditional *t* suite of tests for identifying differentially expressed genes in a DNA microarray experiment with little replication. *Statistics in Biopharmaceutical Research, 1,* 26–38.

Amaratunga, D., & Cabrera, J. (2016). High-dimensional data. Invited review. *Journal of the National Science Foundation of Sri Lanka, 44,* 3–9.

Amaratunga, D., Cabrera, J., Cherkas, Y., Lee, Y. S. (2012). Ensemble classifiers. In D. Fourdrinier, É. Marchand, & A. L. Rukhin (Eds.), *IMS collection volume 8, contemporary developments in Bayesian analysis and statistical decision theory: A Festschrift for William E. Strawderman*.

Amaratunga, D., Cabrera, J., & Lee, Y. S. (2008a). Enriched random forests. *Bioinformatics, 24,* 2010–2014.

Amaratunga, D., Cabrera, J., & Kovtun, V. (2008b). Microarray learning with ABC. *Biostatistics, 9,* 128–136.

Amaratunga, D., Cabrera, J., & Shkedy, Z. (2014). *Exploration and analysis of DNA microarray and other high dimensional data*. New York: Wiley.

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. *Annals of Statistics, 29,* 1165–1188.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24,* 123–140.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika, 68,* 589–599.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association, 78,* 316–331.

Efron, B., & Tibshirani, R. (1997). Improvement on cross-validation: the.632+bootstrap method. *Journal of the American Statistical Association, 92,* 548–560.

Fisher, R. A. (1925) Statistical methods for research workers. Edinburgh: Oliver & Boyd.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software,* 1–22.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika, 58,* 453–467.

Kasim, K., Shkedy, Z., Kaiser, S., Hochreiter, S., Talloen, W. (2016). *Applied biclustering methods for big and high-dimensional data using R*. Chapman & Hall / CRC Biostatistics Series.

Moechars, D., et al. (2005). Sialin-deficient mice: A novel animal model for infantile free sialic acid storage disease (ISSD). In *Society for Neuroscience 35th Annual Meeting*.

Pavlidis, P., et al. (2004). Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochemistry Research, 29,* 1213–1222.

Raghavan, N., Amaratunga, D., Cabrera, J., Nie, A., Jie, Q., & McMillian, M. (2006). On methods for gene function scoring as a means of facilitating the interpretation of microarray results. *Journal of Computational Biology, 13,* 798–809.

Raghavan, N., De Bondt, A., Talloen, W., Moechars, D., Göhlmann, H., & Amaratunga, D. (2007). The high-level similarity of some disparate gene expression measures. *Bioinformatics, 23,* 3032–3038.

Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology, 3*, Article 3.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, B, 36,* 111–147.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, B, 64,* 479–498.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, B, 58,* 267–288.

Tryputsen, V., Cabrera, J., De Bondt, A., & Amaratunga, D. (2014). Using Fisher's method to identify enriched gene sets. *Statistics in Biopharmaceutical Research, 6,* 154–162.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Tukey, J. W. (1980). *Lecture notes for statistics 411. Princeton University* (Unpublished).

Van Acker, N., Verheijen, F., Goris, I., Daneels, G., Schot, R., Verbeek, E., et al. (2017). Progressive leukoencephalopathy impairs neurobehavioral development in sialin-deficient mice. *Experimental Neurology, 291*, 106–119.

Wouters, L., Goehlmann, H., Bijnens, L., Kass, S. U., Molenberghs, G., & Lewi, P. J. (2003). Graphical exploration of gene expression data: A comparative study of three multivariate methods. *Biometrics, 59,* 1131–1140.

Chapter 5
Synergy or Additivity—The Importance of Defining the Primary Endpoint and the Approach to Its Statistical Analysis—A Case Study

Bruce E. Rodda

5.1 Introduction

When a patient is prescribed two different therapeutic moieties for an indication, these entities may be taken either concurrently or in combination. In either case, it is anticipated that co-administration of two agents will result in efficacy and/or safety that is superior to that expected from either component if administered independently. The therapeutic outcome from such a co-administration should benefit from both products, and the contribution of each component should be measurable.

There are several reasons for administering two medications as a single dosage form. These include convenience for the patient, improved compliance, and lower cost, among others. There are also advantages for the sponsor, including manufacturing efficiencies, improved market share, and intellectual property benefits. The latter two benefits may be enhanced if it can be demonstrated that the clinical outcome associated with the administration of the combination is associated with a synergistic effect of the two components. For this reason, demonstrating synergy of two therapeutic entities can be a very important objective in the clinical development program of a combination product.

5.2 Definition

In a combination product, it is important that each entity contribute individually to the overall response. A patient's response to the combination treatment that exceeds the response expected from the simple sum of the individual effects can provide an

B. E. Rodda (✉)
Strategic Statistical Consulting LLC, University of Texas School of Public Health, Austin, TX, USA
e-mail: Bruce_Rodda@msn.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_5

additional therapeutic benefit to the patient and a potential marketing advantage for the sponsor.

The standard definition of synergy of two pharmacologic agents is characterized by the following quotation. "A synergistic effect is one in which the combined effect of two chemicals is greater than the sum of the effects of each agent given alone." (Hardman et al. 1996). In contrast, "[a]n additive effect describes the combined effect of two chemicals that is equal to the sum of the effects of each agent given alone." (Hardman et al. 1996). These definitions are critical to the discussion at hand and are those that have been traditionally used in the evaluation of potential synergy between therapeutic agents.

5.3 Background

While the data used in this chapter are hypothetical, they reflect an actual set of studies that was designed and conducted by a major pharmaceutical company. The example is unique in that the original sponsor used a set of well-designed studies in support of a new drug application that supported additivity of two component medicines, and several years later a different company (Company X) used the identical set of clinical trials and their results to support a patent application that claimed synergy between the two components. The two companies had two different objectives and analyzed the results of these studies in two very different ways to address those different objectives.

The audiences of these two contrasting objectives were also different. The original evaluation was submitted to the US Food and Drug Administration (FDA) by the sponsor in support of a new drug application. Several years later, an independent analysis of the same set of studies was submitted to the US Patent and Trademark Office (USPTO) by Company X in support of a patent application.

The objective of this chapter is to demonstrate that the ability to characterize, and to claim the existence of, a synergistic relationship between two agents depends to a critical degree on how "synergy" is defined and how the data are analyzed.

While the studies and data presented in this chapter are hypothetical, they reflect an actual case and the descriptions of the studies have been extensively modified to preclude associating these results with specific companies, products, or outcomes of the actual studies. Only enough information has been retained to facilitate presentation of the chapter's thesis.

5.4 The Sponsor's Case

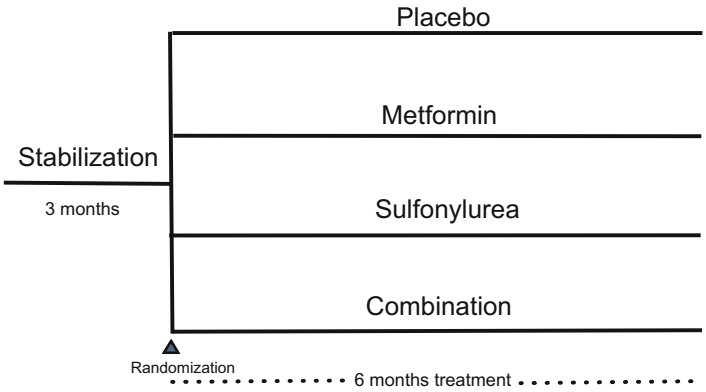
The use of combination products for the treatment of type-2 diabetes has been common for several years, and the example used in this chapter is based on a development

plan for a combination of metformin and a sulfonylurea for the treatment of this disease.

The goal of the sponsor in this example was to develop a combination of metformin and a sulfonylurea in this indication, using the reduction of hemoglobin A1c (hereafter A1c) following six months of treatment as the primary outcome variable. Their interest was in providing a product with the benefits of both agents and did not anticipate any synergistic or potentiative activity of the two component medications, although their clinical development plan allowed evaluation of a potential synergistic effect between the two agents. The original submission by the sponsor to the FDA requested that the combination be approved as safe and effective for the treatment of this indication, but there was no request for a claim of synergy between the two agents.

As the basis of their new drug application (NDA), the sponsor conducted five clinical trials in support of the combination product. These clinical trials will form the basis for the discussion of the case presented in this chapter.

The designs of four of these studies were standard 2×2 designs in which the four treatments were placebo, metformin, sulfonylurea, and the combination of metformin and sulfonylurea. These designs are presented in the following schematic.



The treatment layout for the first four studies was:

Studies 1–4

	Placebo sulfonylurea	Sulfonylurea
Placebo metformin	Placebo	Sulfonylurea
Metformin	Metformin	Combination

Two doses of metformin (low/high) and three doses of sulfonylurea (low/med/high) were included in the clinical development plan. The 2×2 designs used different combinations of doses, and the results of these studies will be discussed later in this chapter.

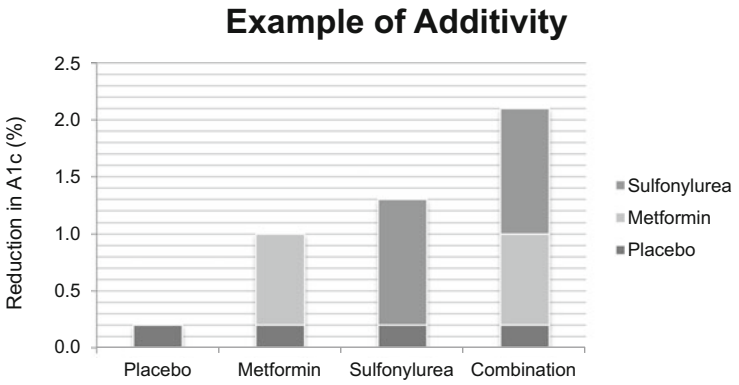
The fifth study was a 3×4 design in which two doses of metformin and three doses of sulfonylurea were also included, resulting in 12 treatment groups (M-lo, M-hi, S-lo, S-med, S-hi, Placebo, M-lo/S-lo, M-lo/S-med, M-lo/S-hi, M-hi/S-lo, M-hi/S-med, and M-hi/S-hi).

Study 5

	Placebo sulfonylurea	Sulfonylurea (low)	Sulfonylurea (med)	Sulfonylurea (high)
Placebo metformin	placebo	S(lo)	S(med)	S(hi)
Metformin (low)	M(lo)	M(lo)/S(lo)	M(lo)/S(med)	M(lo)/S(hi)
Metformin (high)	M(hi)	M(hi)/S(lo)	M(hi)/S(med)	M(hi)/S(hi)

These five studies followed standard designs for the evaluation of potential synergy (metformin by sulfonylurea interaction) and would have been capable of identifying synergy, if it existed.

The concept of pharmacologic additivity can be appreciated by the graphic example below. Consider the design of the four simpler studies, each comprising four treatment groups. Assume that the placebo treatment is associated with a 0.2% reduction in average estimated A1c over the six-month observation period; the patients in the metformin group have a 0.8% reduction in A1c: and the patients in the sulfonylurea group have a 1.1% reduction in A1c. These reductions are generally consistent with the treatment effects observed in the studies under consideration.



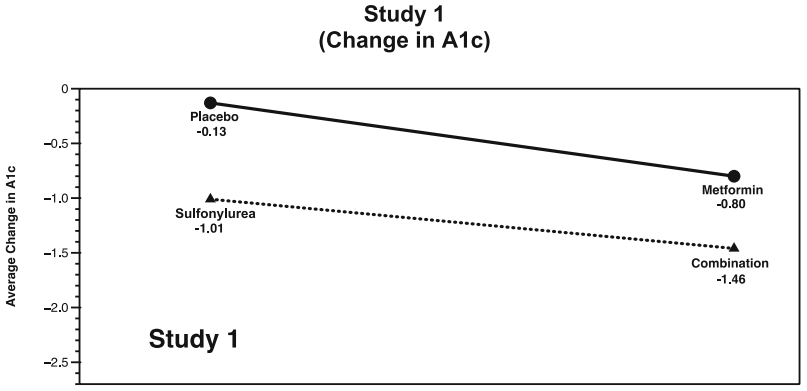
Note that the effect of every treatment contains a response due to placebo, and we must subtract this effect from the observed treatment response to obtain an estimate of the net (or pharmacologic) effect for each treatment. In this example, the net (pharmacologic) effect of metformin is 0.8% (1.0%–0.2%) and that of sulfonylurea is 1.1% (1.3%–0.2%). The figure above demonstrates that the observed effect for each treatment is the total of the net pharmacologic effects and the placebo effect.

To relate this concept to the metformin and sulfonylurea combination treatment, refer to the column on the far right in the chart above. If metformin and sulfonylurea

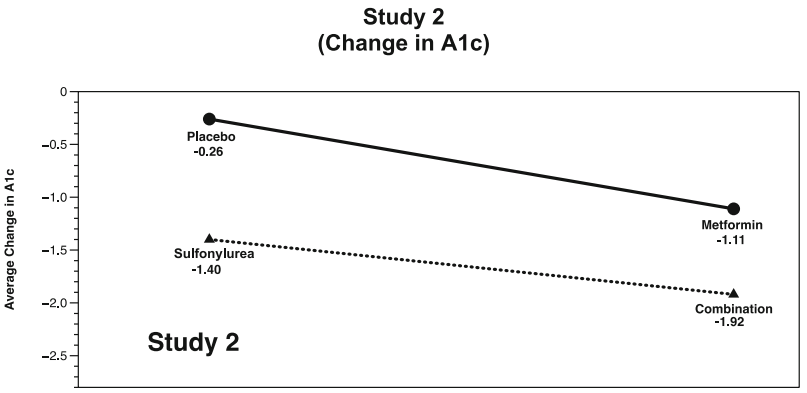
are truly additive in their pharmacologic effects at these doses, the expected response (reduction in A1c) associated with the combination of metformin and sulfonylurea (the sum of the effects taken independently) would be the sum of the individual effects, i.e., the placebo effect which is common to every treatment (0.2%) plus the metformin effect (0.8%) plus the sulfonylurea effect (1.1%) which yields an expected reduction of 2.1% for the metformin and sulfonylurea combination treatment.

Synergy would only occur if "the combined effect of two components is greater than the sum of the effects of each agent given alone." (Hardman et al. 1996). Thus, for the combination of metformin and sulfonylurea to be truly synergistic in the example above, the average reduction in A1c associated with the combination would need to exceed 2.1% by an amount greater than would be expected by chance. This would be characterized by a large metformin by sulfonylurea interaction effect in the various statistical analyses.

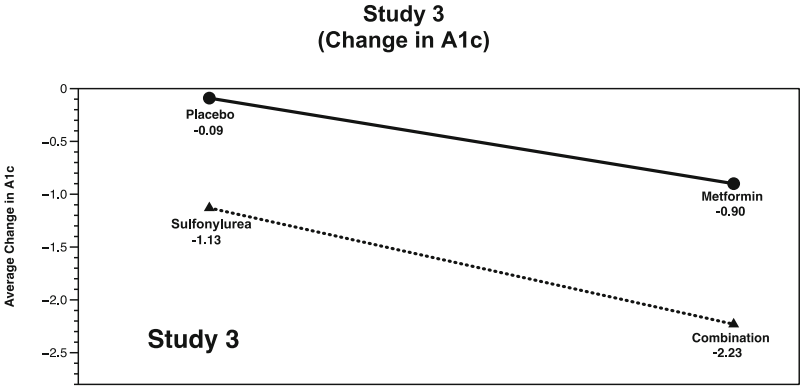
Although results are usually presented later in a report, it is important to present summaries of these five trials at this point to form the basis for the central argument of this chapter. In addition, the thesis of this chapter is not related to sample size or missing data. Therefore, only the average responses to the treatments in the individual studies will be presented.



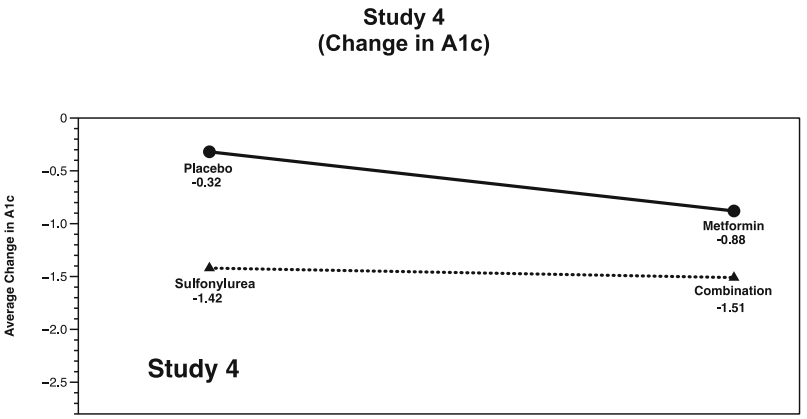
In the first study, the test for a metformin by sulfonylurea interaction did not suggest the presence of any interaction (synergy). The expected effect of the combination in the absence of synergy was −1.68 (−0.80 + (−1.01) + 0.13), very similar to the observed difference of −1.46.



There was no suggestion of synergy in Study 2 either, the expected effect of the combination in the absence of synergy being −2.23 (−1.11 + (−1.40) + 0.28), also similar to the observed difference of −1.92.

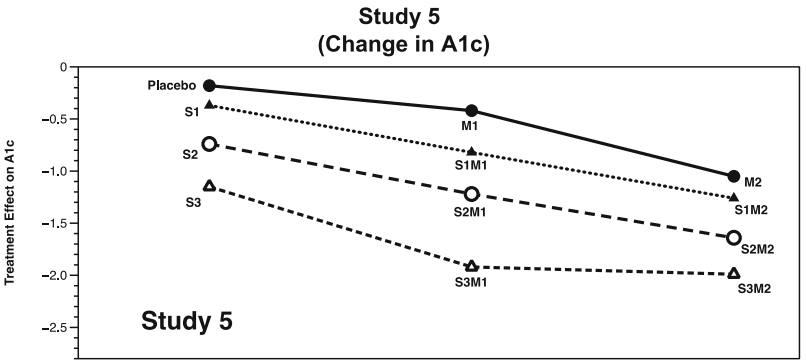


The statistical analysis of Study 3 indicated additive effects of metformin and sulfonylurea at these doses. The estimated effect if the effects were additive was −1.94 (−0.90 + (−1.13) + 0.09), comparable to the observed value of −2.23.



In the fourth study, the effect of the combination of the two treatments was similar to the estimated effect of sulfonylurea alone. The statistical analysis was supportive of both metformin and sulfonylurea effects at these doses, but no synergy was implied. Note that these first four studies offer no suggestion of any synergistic effect of the two agents over the dose ranges explored. The parallel lines in the graphs visually support additivity of the agents.

The fifth and final study explored a range of doses for both entities as previously described and also provided no evidence of a synergistic effect of the two drugs. The results of this study are presented in the figure below.



The table below presents the estimates of the net treatment effect (actual observation minus placebo effect) for Study 5. In addition, the numbers in parentheses are the estimates of the combination treatment effects that would be expected if there was complete additivity. These numbers, the statistical analysis, and the graphic above provide no evidence of synergy of any kind in this study.

Study 5 (Change from placebo in A1c) Sulfonylurea

		Placebo	Low	Medium	High
Metformin	Placebo	–	−0.19	−0.56	−0.97
	Low	−0.24	−0.64 (−0.43)	−1.04 (−0.80)	−1.74 (−1.21)
	High	−0.87	−1.08 (1.06)	−1.46 (−1.43)	−1.81 (−1.84)

The results of the five studies provide ten opportunities for synergy to be observed, if it existed. Each of the four 2×2 studies provides a single estimate and the 3×3 study provides six estimates of possible synergy. Each of these five studies was independently analyzed by the original sponsor using general linear model theory and included factors for metformin, sulfonylurea, and their interaction. The estimated treatment effects are summarized in the table below

Summary of Patient Data

Average treatment effects submitted to the FDA in consideration of synergy					
Study M/S dose	Metformin	Sulfonylurea	Combination	Expected effect if additive	Difference implying synergy
1-M/H	−0.67	−0.87	−1.33	−1.54	0.21 (N)
2-H/H	−0.85	−1.14	−1.66	−1.99	0.33 (N)
3-M/L	−0.81	−1.04	−2.14	−1.85	−0.29 (Y)
4-L/L	−0.56	−1.10	−1.19	−1.66	0.47 (N)
5-L/L	−0.24	−0.19	−0.64	−0.43	0.21 (N)
5-M/L	−0.24	−0.56	−1.04	−0.90	−0.24 (Y)
5-H/L	−0.24	−0.97	−1.74	−1.21	−0.53 (Y)
5-L/H	−0.87	−0.19	−1.08	−1.06	−0.02 (Y)
5-M/H	−0.87	−0.56	−1.46	−1.43	−0.03 (Y)
5-H/H	−0.87	−0.97	−1.81	−1.84	0.03 (N)

The final column in the table above presents the difference between the observed effect of each of the combinations and the effect expected if there was an additive relationship between the two components. In this column, a positive number indicates an effect less than additivity and a negative number indicates potential positive synergy. The statistical analyses of the individual studies resulted in p values exceeding 0.20 for all interactions. Note that from a purely directional point of view, one-half the observed differences between the actual combination effect and hypothesized additivity was negative and one-half was positive—exactly what would be expected if the two components were additive. In addition, in each analysis there was firm statistical support for the clinical effect of each component treatment.

While it would be inappropriate to perform any statistical tests on the data in the final column of the table for a variety of reasons, there clearly is no suggestion from these studies that metformin and sulfonylurea are synergistic in reducing A1c after six months of treatment.

The conclusions that are suggested by these studies are that each of the two component medicines is associated with reductions in A1c and that the combination is also associated with a reduction in A1c that is consistent with additive contributions of both components. This was the conclusion of both the sponsor and the FDA. The combination product was approved for the treatment of type 2 diabetes; no request was made for a statement of synergy in the labeling and none was granted.

Since there did not appear to be any evidence of synergy associated with the clinical effects of these two compounds, no patent claiming synergy was filed by the sponsor with the US Patent and Trademark Office (USPTO).

5.5 Company X's Case

As the term of the original exclusivity neared expiration, interest in a generic version of the combination was pursued by other firms. One of these companies (Company X) determined that a patent that supported synergy of the two components in this indication would provide them with a unique position in the market, provide a potential marketing advantage, and thus improve sales. For these reasons, they decided to file a patent application seeking a claim of synergy between the two agents. The foundation for their filing included the identical studies and data that the original sponsor had submitted to the FDA in support of the approved NDA. However, since the original analyses of these studies did not support synergy, Company X needed a different approach to support the patent application of potential synergy.

It should be pointed out that in contrast to an approved new drug application, a patent does not assure that a product provides the effect cited in the patent. The patent provides a claim of intellectual property, but that claim may not necessarily be related to the product's actual performance in humans—even if the product's performance in humans is known. In addition, and unlike submissions to the FDA, submissions to the USPTO can be based on selective datasets and analyses. There is no requirement for full disclosure regarding all knowledge available regarding the product. Selective data supporting a patent application may be submitted to the USPTO without providing information that may conflict with the intended claim for clinical use, and there is no requirement for consistency between submissions to the USPTO and the FDA.

The analytic approach that was used by Company X to support their claim of synergy differed greatly from that traditionally used to evaluate synergy (as used by the original sponsor) and was of questionable statistical validity. However, Company X maintained in their patent application that their definition of synergy was the same as the original sponsor, i.e., "a synergistic effect is one in which the combined effect

of two chemicals is greater than the sum of the effect of each agent given alone." (Hardman et al. 1996).

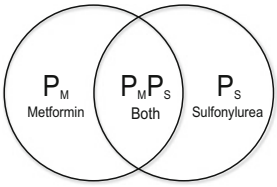
Despite the assertion by Company X that their definition of synergy was the same as that of the original sponsor, the analyses supporting their contention of synergy were completely different from the analyses submitted to the FDA by the original sponsor, whose original analyses were inconsistent with a synergistic effect.

The approach used by Company X was based on the concept of "responder." This technique required that each patient be characterized as a "responder" or "non-responder" to their assigned treatment, thus effectively ignoring the degree of response for each patient beyond that required to determine whether a patient was a responder.

A "responder" in this case was defined by Company X as a patient whose A1c was less than 6.0% or who experienced a reduction of at least 1.0% after six months of treatment. The percentage of responders in each treatment group was then computed within each study according to this definition and was then used as the basic datum for comparison.

In contrast to the classic definition of synergy discussed previously, the only approach to support synergy submitted to the USPTO by Company X was based on the following methodology. First, a "placebo-adjusted" proportion of patients who responded in each active treatment group was calculated. Company X defined the "placebo-adjusted" effects for each study to be the percentage of responders in each active treatment group minus the percent responders in the placebo group in that study. While this clearly could be problematic (the difference could be less than 0.0), their strategy for the USPTO submission was based on this approach. Fortunately for Company X, none of these "placebo-adjusted" probabilities was negative or the irrationality of their approach would have been obvious to the USPTO.

The rationale for this approach can be appreciated by viewing the Venn diagram below and the subsequent explanation.



In the schematic above, P_M represents the "adjusted" probability of response for the metformin groups and P_S represents the "adjusted" probability of response for the sulfonylurea groups.

This approach has been validly used (without the "adjustment") to determine synergy between two herbicides (Hewlett and Plackett 1979). Consider the following example of two herbicides (say M and S) designed to kill a given weed. Assume that M is effective in killing a proportion (P_M) of the plot if administered alone and S is effective in killing a proportion (P_S) of the plot when administered alone. If the two

herbicides were applied together and acted independently, the proportion of the plot being killed would be $P_M + P_S - P_M P_S$ based on conventional probability theory.

Thus, if there was no synergy, the expected proportion of the plot dying with combination treatment would be $P_M + P_S - P_M P_S$. If the observed proportion of the plot that was killed exceeded $P_M + P_S - P_M P_S$, there would be evidence of a synergistic effect of the two herbicides. While this approach makes sense in the context of a situation where the totality of effect cannot exceed unity (the entire plot being killed), it does not make sense in the context of demonstrating synergy of drugs with measurable effects. In addition, the definition of synergy in this context is different than the additive definition in which the combined effect of two chemicals is greater than the sum of the effects of each agent given alone.

This rationale for demonstrating synergy fails on both logical and statistical grounds when there is no upper bound on the treatment effect (unlike death).

To demonstrate the fallacies of this approach, the proportion of responders in each study according to Company X's definition is summarized in the following table.

Summary of Patent Data

Entries are percent of "responders" for each treatment

Treatment	Protocol 1	Protocol 2	Protocol 3	Protocol 4	Protocol 5
Placebo	15.8	24.4	18.3	17.7	34.7
Low M	–	–	–	20.0	44.3
Med M	–	38.3	37.7	–	40.0
High M	53.3	–	–	–	57.1
Low S.	–	–	40.5	44.6	46.5
High S.	67.5	50.0	–	–	58.9
Lo M/Lo S.	–	–	–	37.9	69.4
Lo M/Hi S.	–	–	–	–	78.1
Med M/Lo S.	–	–	61.5	–	60.6
Med M/Hi S.	–	66.4	–		80.9
Hi M/Lo S.		–	–	–	68.6
Hi M/Hi S.	87.0	–	–	–	81.2

The rationale for determination of the existence of synergy in each study was based on the logic presented above using the adjusted proportion of responders in each active treatment group. As discussed previously, this adjustment was made by subtracting the proportion of placebo responders from the proportion of responders in each treatment group to achieve a "placebo-adjusted" response rate. For example, in Protocol 1 the adjusted response rate (%) for the high dose metformin treatment would be $53.3 - 15.8 = 37.5$, that for the high dose of sulfonylurea would be $67.5 - 15.8 = 51.7$, and that for the combination would be $87.0 - 15.8 = 71.2$.

The evaluation of synergy was then made by determining whether the observed "placebo-adjusted" proportion of responders in the combination group exceeded that which would be expected using the $P_M + P_S - P_M P_S$ definition of additivity. In this

case, $P_M + P_S - P_M P_S$ would be $0.375 + 0.517 - (0.375) * (0.517) = 0.698$ or 69.8%. Since the observed adjusted response rate for the combination in this study was 71.2%, there is evidence of synergy by this approach. These calculations have been performed for each of the various studies and combinations and are presented below.

Summary of Patent Data

Entries are "placebo-adjusted" response probabilities (percentages) submitted to the USPTO in support of synergy

Study M/S dose	% Metformin (M)	% Sulfonylurea (S)	M + G – M * S	Observed combination	Difference implying synergy
1-M/H	37.5	51.7	69.8	71.2	1.4 (Y)
2-H/H	13.9	25.6	35.9	42.0	6.1 (Y)
3-M/L	19.4	22.2	37.3	43.2	5.9 (Y)
4-L/L	2.3	26.9	28.6	20.2	−8.4 (N)
5-L/L	9.6	11.8	20.3	34.7	14.4 (Y)
5-M/L	5.3	11.8	16.5	25.9	9.4 (Y)
5-H/L	22.4	11.8	31.6	33.9	2.3 (Y)
5-L/H	9.6	24.2	31.5	43.5	11.9 (Y)
5-M/H	5.3	24.2	28.2	46.2	18.0 (Y)
5-H/H	22.4	24.2	41.2	46.5	5.3 (Y)

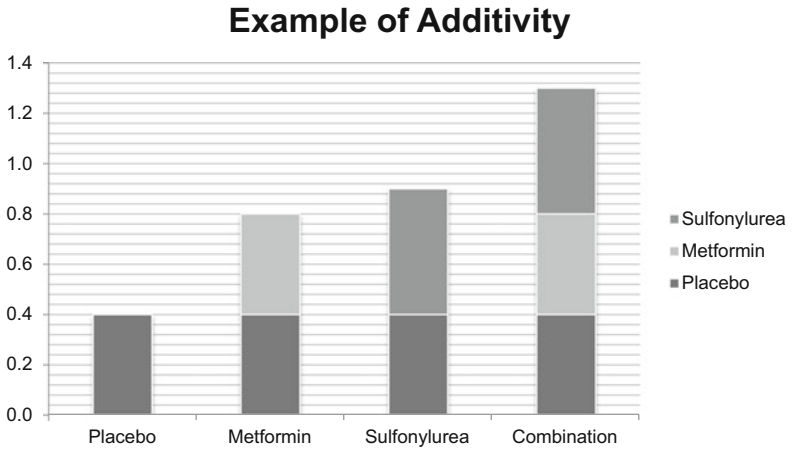
The table above was the general summary submitted to the USPTO to support a claim of synergy. The position of Company X was that there were ten treatment combinations among the five studies that could be used to determine the existence of synergy, and nine supported synergy. They also cited that a binomial test of the ten results was associated with a p value of < 0.01. None of the A1c data from the five studies that were submitted to the FDA were submitted to the USPTO. Based on this limited information, but absent any of the analyses and summaries submitted to the FDA, a patent for synergy was granted by the USPTO.

5.6 Discussion

The hypothetical example above is based on an actual case and clearly distinguishes between two approaches to the evaluation of potential synergy between two agents. The first approach uses the factorial nature of the study designs and provides estimates of potential interaction effects, which in this case was not consistent with synergy. The second approach, using consistency with probabilistic independence ($P_M + P_S - P_M P_S$) as an index of synergy, has several elements that make it inappropriate for use in a clinical trial situation.

The first drawback of the second approach is that it requires a definition of "responder." When trying to characterize synergy of herbicides, the responder definition is clear—death, and the proportion of the plot dying is a reasonable metric. The logic in the $P_M + P_S - P_M P_S$ definition of synergy in this case is supportable.

However, the definition of "responder" using this approach in a clinical trial has a significant impact, even if there is no "adjustment" for placebo response and using the fundamental concept manifest in the classical definition of synergy. For example, suppose (but being extreme to make the point) that all placebo patients have a 0.4% reduction in A1c; all metformin patients have a 0.8% reduction; and all patients in the sulfonylurea group have a 0.9% reduction. If the effects are additive, we would expect to see an average reduction of 1.3% in the combination group (0.4%P + 0.4%M + 0.5%S) as in a previous schematic and represented below.



Recall that one of the requirements to be considered a responder is that a patient must experience a reduction of at least 1.0% in A1c over the 6 months of study. In this example and by this definition, none of the placebo patients will be responders; none of the metformin patients will be responders; and none of the sulfonylurea patients will be responders; but 100% of the combination patients will be responders. Using Company X's logic and statistical approach, synergy will be claimed where none exists.

Suppose we make a minor change in the definition of responder and reduce the required reduction in A1c to 0.9% instead of 1.0%. Assume that all placebo patients still have a 0.4% reduction; all metformin patients have a 0.8% reduction; and the patients in the sulfonylurea group still have a 0.9% reduction. If the effects are additive, we would still expect to see an average reduction of 1.3% in the combination group (0.4%P + 0.4%M + 0.5%S).

Using this revised definition of responder, where the only change from the previous definition was a reduction in the requirement from a minimum change of 1.0% to 0.9%, none of the placebo patients will be responders; none of the metformin patients

will be responders; but 100% of the sulfonylurea patients and 100% of the combination patients will be responders. Since the sulfonylurea and combination patients have the same proportion of responders (100% in this example), demonstrating synergy would be impossible using this approach—even if it existed.

Consider a variation of the previous example where the proportion of responders is the same for all single entity treatments as in the previous paragraph, but the average response in the combination group is no longer 1.3%, but is now 2.0%—real synergy. It is still impossible to demonstrate synergy using the "responder" approach since all patients in the sulfonylurea and combination groups were "responders." Here we have clear evidence that there is synergy with respect to the average reduction in A1c, but by using the "responder" approach, it is impossible to detect.

The above discussion identifies several points that make using the proportion of responders to characterize synergy inappropriate when a measureable endpoint is used as the primary efficacy variable.

- The actual degree of response is not considered.
- Synergy can be inferred where it does not exist.
- Synergy may appear to be absent when it truly exists.
- Either of these results may be created by selectively defining the term, "responder."
- Bias could be easily introduced if the definition of responder is selected with knowledge of the study results.

There were other problematic issues with using a responder approach for determining the existence of synergy in this specific case.

- Each treatment combination contributed only one piece of information for the evaluation of synergy.
- The sample sizes of the various studies and the number of patients receiving each of the various combinations were not considered in the evaluation.
- All comparisons using the "responder" approach were based on an assumption that the comparisons were independent, which is clearly not tenable.
- Six of the ten comparisons in support of the patent came from the same study with the same placebo contributing to each comparison.
- Each individual treatment contributed to more than one comparison.
- The placebo-adjusted response rates for active treatments were computed by subtracting the proportion of placebo responders in each study from the proportion of responders in the active treatment groups to obtain a "placebo-adjusted response probability" associated with each active treatment.
- While the theory purporting to support this approach is tenuous at best, it is very possible for the proportion of responding placebo patients to exceed the proportion of patients responding to an active agent. Subtracting the placebo proportion from the active proportion would yield a negative probability—a nonsensical result.

In the hypothetical case described in this chapter, the clinical studies forming the basis for both a new drug application by the original sponsor and a patent application by Company X were identical. However, the analytic approaches used to demonstrate

synergy were very different in the two cases. In the original approach submitted to the FDA, the conventional definition of synergy and general linear model theory was used to evaluate whether synergy existed. This approach used all the data in the studies and determined that no synergy existed, defining synergy as an effect of the combination that exceeds the sum of the effects of the individual components. A claim for synergy was not requested by the sponsor, nor granted by the FDA. The FDA approved the combination as being efficacious, but did not consider the data as supportive of synergy.

An alternative approach to determining synergy was chosen by Company X to support a patent application claiming synergy between the two component medicines. Characterizing each patient as a responder (or non-responder), a probabilistic definition of synergy and its evaluation was used. This approach is quite appropriate if the endpoint of interest is unique and dichotomous. While the definition of synergy in this case is not additive as when a measured outcome is primary, the $P_M + P_S - P_M P_S$ definition is a logical approach to demonstrating synergy by a different definition. In this case, it is critical that the endpoint be unique, since the outcome can be manipulated by selective definition of "response."

The hypothetical example presented in this chapter was based on an actual case and demonstrates the importance of defining the anticipated outcome variable as an integral component of the objective of a clinical development plan (or protocol). Different definitions of outcome variables (endpoints) that are intended to evaluate the same objective can result in different analyses and conflicting conclusions.

References

Hardman, J. G., et al. (1996). *The pharmacological basis of therapeutics*. New York: McGraw-Hill.
Hewlett, P. S., & Plackett, R. L. (1979). *An introduction to the interpretation of quantal responses in biology*. London, UK: Edward Arnold.

Chapter 6
Recycling of Significance Levels in Testing Multiple Hypotheses of Confirmatory Clinical Trials

Mohammad Huque, Sirisha Mushti and Mohamed Alosh

6.1 Introduction

For the demonstration of benefits of a new treatment, confirmatory clinical trials generally include multiple hypotheses and classify them into primary and secondary types and sometimes also into other lower types, such as tertiary, supportive, and exploratory. These trials normally define primary and secondary hypotheses in terms of the primary and secondary endpoints of the trial and frequently assign weights to them in testing based on their clinical importance and power considerations. Thus, the sets or families of hypotheses in confirmatory trials may follow a hierarchically ordered structure, with the primary family holding a special status, so that if the trial wins for one or more of its hypotheses then one can characterize clinically relevant benefits of the study treatment. The role of the secondary hypotheses in the trial is generally to demonstrate additional benefits of the study treatment. O'Neill (1997) and ICH (E-9, 1998) discussed the importance of structuring study endpoints and hypotheses into the primary and secondary types that best reflect the objectives of clinical studies.

The last two decades have witnessed several innovative statistical procedures that account for the above hierarchical structure in testing of multiple hypotheses. Some of these methods recycle the significance level of a successfully rejected hypothesis to other hypotheses within the same family (e.g., the primary family) and to

M. Huque (✉)
Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, USA
e-mail: huque.stat@gmail.com

S. Mushti
Division of Biometrics V, Office of Biostatistics, OTS, CDER, FDA, Silver Spring, MD, USA

M. Alosh
Division of Biometrics III, Office of Biostatistics, OTS, CDER, FDA, Silver Spring, MD, USA

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_6

hypotheses in other families according to an "α-propagation rule," which shows how the significance level of a rejected hypothesis distributes to other hypotheses to be tested. Certain methods take a more a general approach. For example, the graphical approach of Bretz et al. (2009), reviewed in Sect. 6.5, considers hypotheses of a trial individually instead of structuring them into families; the directed edges of the graph which link different hypotheses are used to represent the hierarchical and logical restrictions that may be applicable when testing a number of hypotheses.

The idea of hierarchical testing for a family of multiple hypotheses has been around for some time. For example, Williams (1971) used such a technique to test for minimum effective dose using his newly introduced trend test. Maurer et al. (1995) and Bauer et al. (1998) have used the hierarchical testing approach; Maurer et al. in their article gave the first "modern" proof for it using closed testing which we address in Sects. 6.3 and 6.4. However, Westfall and Krishen (2001) first used the term "gatekeeping" in the context of testing of hierarchically ordered families $F_1, \ldots, F_s$ of hypotheses, with $s \geq 2$. They showed that the Type I error rate is controlled when the family F_{i+1} assigned the full trial level α after all hypotheses in the preceding family F_i are first rejected, for $1 \leq i \leq s-1$. This testing approach that allows recycling of entire α from a family to the next later became known as the *serial gatekeeping strategy*. Dmitrienko et al. (2003) later proposed another test strategy in which the family F_{i+1} can be tested if at least one hypothesis in the family F_i is rejected, for $1 \leq i \leq s-1$. They call this test strategy as the *parallel gatekeeping strategy.* Both the serial and parallel gatekeeping strategies guarantee strong control of the familywise error rate (FWER) as defined in Hochberg and Tamhane (1987). These above articles along with articles by Bauer (1991), Koch and Gansky (1996), and Hsu and Berger (1999) triggered considerable research activity on this topic with subsequent contributions by Dmitrienko et al. (2006), Dmitrienko and Tamhane (2007), Hommel et al. (2007), Dmitrienko and Tamhane (2009), and many others.

Further, Dmitrienko et al. (2008) discussed that the Holm (1979) and Hochberg (1988) procedures do not allow recycling of α from the primary family to the secondary family unless all hypotheses in the primary family are first rejected. This is the case, for example, in a two-arm trial which compares a treatment to control on multiple primary endpoints using either the Holm or the Hochberg procedure. They proposed modifications of these procedures by truncating them. Their truncated versions of the Holm and Hochberg procedures do allow recycling of α from the primary family to the secondary family of hypotheses when at least one of the hypotheses of the primary family is rejected. Dmitrienko et al. (2013) and Huque et al. (2013) discussed gatekeeping strategies and truncated Holm and Hochberg procedures for clinical trial applications.

In addition, Wiens (2003) and Wiens and Dmitrienko (2005) proposed an extension of the fixed-sequence method calling it the *fallback* method. This method allows testing a hypothesis in the sequence even when a preceding hypothesis in the sequence is not rejected. Huque and Alosh (2008) later extended the fallback method to consider the correlation between the test statistics when the joint distribution of the test statistics is specified; this method is now known as the *parametric fallback* method. Further, a key article by Hommel et al. (2007) showed that several standard test pro-

cedures, such as Holm, weighted Bonferroni–Holm, fixed-sequence, and fallback methods, and certain gatekeeping procedures immediately follow from shortcuts to the closed testing principle of Marcus et al. (1976) when using Bonferroni adjustments.

Li and Mehrotra (2008), on the other hand, introduced a concept of adapting the significance level for testing secondary hypotheses based on the finding of the primary hypotheses tests. Alosh and Huque (2009), likewise, on considering ideas from Song and Chi (2007), introduced the notion of *consistency* in testing for an effect in the overall population and in a specific subgroup. These authors later extended this consistency concept to other situations, including the description of a consistency-adjusted strategy for accommodating an underpowered primary endpoint; see Alosh and Huque (2010) and Huque and Alosh (2012). Research related to consistency in testing multiple hypotheses continues; see, for example, Li (2013), Alosh et al. (2014), and Rauch et al. (2014). Huque et al. (2012) used some of these methods for testing multiple hypotheses of composite endpoint trials. Also, Hung and Wang (2009, 2010) proposed some controversial multiple testing problems.

Additionally, Bretz et al. (2009) and Burman et al. (2009) independently introduced a nifty graphical framework for creating and visualizing test strategies for common multiple test problems. Bretz et al. (2011a, 2011b) provided further understandings about the use of this approach for clinical trial applications. This approach, though similar in concept to the general gatekeeping approach, allows recycling of significance level of a successfully rejected hypothesis to other hypotheses according to an alpha-propagation algorithm in a way that preserves the FWER for the trial in the strong sense. Its appeal notably is in the graphical visualizations of how alpha-allocations are initially made to primary and secondary hypotheses, and after a hypothesis is rejected, how these alpha-allocations modify and shift to other remaining hypotheses for subsequent tests. In addition, its easy use allows evaluation of multiple design options for tailoring the trial design down to a level that can have a greater chance of success. Free software package gMCP in *R* is now available for using this method; see Bretz et al. (2011b) on how to download and use this package.

However, on examining the graphical approach with the work by Hommel et al. (2007), one finds that this approach in its simplest form is a by-product of the shortcut to the closed testing approach when applying weighted Bonferroni tests to the intersection hypotheses, as shown in Bretz et al. (2009). In the graphical approach, weights for the weighted Bonferroni tests are updated after each hypothesis rejection through an algorithm so that the *consonance condition* holds, thus allowing shortcuts to the closed testing approach, which can then be represented in a graphical form. This shortcut closed testing approach and its graphical representation also extend to the weighted parametric and Simes tests of hypotheses when the joint distribution of the test statistics can be fully or partially specified; see Bretz et al. (2011b).

The aforementioned contributions, and others not mentioned here, remind us that the last two decades have witnessed the introduction of many important statistical methods for addressing multiplicity problems of clinical trials. Many of them are now used in designing modern clinical trials with multiple objectives. Particularly, methods related to recycling of significance levels in testing multiple hypotheses,

when these hypotheses follow a hierarchical structure, are of significant importance for clinical trial applications, as they reduce the degree of multiplicity and maximize power for the primary hypotheses test of the trial. However, some of these methods, as they are fairly new, are not written down in books for easy understanding. This chapter is therefore an effort in that direction.

This chapter summarizes concepts and methods for recycling of significance levels in testing multiple hypotheses for the fixed-sample trial designs. Readers interested in allocating recycled significance levels in group sequential procedures when testing multiple hypotheses may refer to Maurer and Bretz (2013), Xi and Tamhane (2015) and Chap. 7 of this book. Section 6.2 introduces concepts regarding recycling of significance levels. Sections 6.3 and 6.4 then introduce closed testing and shortcut closed testing procedures. Section 6.5 then gets into the sequentially rejective graphical procedures and explores their relationship to the shortcut closed testing. Finally, Sect. 6.6 makes some concluding remarks. In this chapter, unless otherwise specified, words used as "null hypothesis" or "hypothesis" are synonymous, including their plural versions "null hypotheses" and "hypotheses." Also, the reported significance levels and p-values are for two-sided tests, unless mentioned otherwise.

6.2 Recycling of Significance Levels in Testing Multiple Hypotheses

The recycling of significance levels follows from the idea that if one rejects H_1 at a significance level α_1, then this α_1 can be recycled such that H_2 can be tested at level $\alpha_2 + \alpha_1 = \alpha$, instead of α_2. One can see this recycling of significance level, for example, in the Holm (1979) procedure for testing two hypotheses, where $\alpha_1 = \alpha_2 = \alpha/2$. Let p_1 and p_2 denote the p-values associated with the tests of the two hypotheses H_1 and H_2, respectively. The Holm procedure first tests the hypothesis associated with the smaller of the two p-values at level $\alpha/2$. If that p-value is less than $\alpha/2$, it rejects this hypothesis and tests the other hypothesis (hypothesis associated with the larger of the two p-values) at the full significance level α. An equivalent graphical procedure of the Holm procedure with two hypotheses shows explicitly how the recycling of the significance levels occurs upon the rejection of a hypothesis (see Fig. 6.1). This graphical procedure initially assigns $\alpha/2$ to H_1 and $\alpha/2$ to H_2 and then tests each of the two hypotheses at level $\alpha/2$. If the procedure rejects one of the two hypotheses, then the significance level $\alpha/2$ of this rejected hypothesis is recycled, so that the other hypothesis is tested at the full significance level α. For example, if the procedure initially rejects H_2 on observing $p_2 < \alpha/2$, then this $\alpha/2$ of H_2 is recycled to test the other hypothesis H_1 at level α, as a result of adding the recycled $\alpha/2$ from H_2 to the original $\alpha/2$ from H_1.

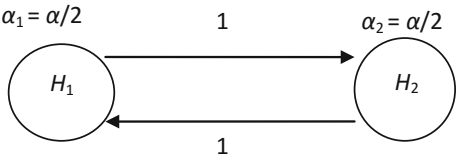


Fig. 6.1 Graphical representation of Holm's test with two hypotheses H_1 and H_2 with $\alpha = 0.05$

6.2.1 *Recycling of Significance Levels in Testing Hierarchically Ordered Hypotheses*

Recycling of significance levels can also be seen in the fixed-sequence test procedure which is often used for testing multiple hypotheses of clinical trials when the hypotheses are hierarchically ordered in pre-specified testing sequence. The first hypothesis is assigned with full significance level α, and the remaining hypotheses are initially assigned with level 0. If the first hypothesis is rejected at level α, then this α is recycled to test the second hypothesis in the sequence, so that the second hypothesis can now also be tested at full level α. If the second hypothesis is also rejected at level α at this stage, then the procedure proceeds to test the third hypothesis in the sequence. This test strategy controls the FWER as long as (1) the testing sequence is specified prospectively, and (2) there is no further testing as soon as a hypothesis is not rejected. The idea behind this test strategy is that a rejection of a hypothesis at a level α allows recycling that significance level to the next hypothesis in the testing sequence. However, the fixed-sequence test procedure stops testing (i.e., no further recycling allowed) as soon as it fails to reject a hypothesis.

In order to justify that the above fixed-sequence test strategy controls the FWER in the strong sense, consider first the simplest case of testing two hypotheses H_1 and H_2. The procedure tests H_1 first at level α, and if H_1 is rejected, then it tests H_2 at the full level α. There are three null hypothesis configurations $H_1 \cap H_2$, $H_1 \cap K_2$, and $K_1 \cap H_2$, where K_1 and K_2 denote the alternative hypotheses of H_1 and H_2, respectively. For the case $K_1 \cap H_2$, the Type I error rate does not exceed α, because the Type I error is committed only in falsely rejecting H_2 and this probability is α by construction. For the case $H_1 \cap K_2$, the Type I error rate does not exceed α for similar reasons. For the case $H_1 \cap H_2$, let R_1 and R_2 denote the rejection regions for testing H_1 and H_2, respectively, so that $\Pr(R_1|H_1) = \alpha$ and $\Pr(R_2|H_2) = \alpha$. Then, $\Pr(R_1 \cup R_2|H_1 \cap H_2) = \Pr(R_1|H_1) = \alpha$, because R_2 is a subset of R_1 since H_2 is tested only after H_1 is first rejected at level α. This completes the proof for the case of testing two hypotheses. The proof for the general case follows from the results of Maurer et al. (1995) and also from the application of the shortcuts to closed testing shown in Hommel et al. (2007).

The appeal for the fixed-sequence testing strategy is that each hypothesis can be tested at the largest possible significance level α. However, its main drawback is that if a hypothesis in the sequence is not rejected, then subsequent hypotheses cannot be rejected even if one or more of them have extremely small p-values. For example, consider a trial which tests hypotheses H_1 and H_2 in a sequence and tests

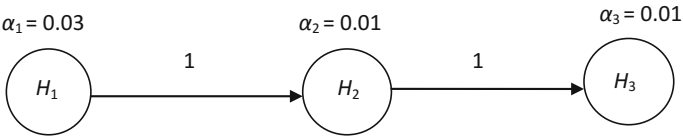


Fig. 6.2 Graphical illustration of the fallback procedure for testing three hypotheses H_1, H_2, and H_3

H_1 first. Suppose that the p-value for H_1 is $p_1 = 0.22$ and the p-value for H_2 is $p_2 = 0.0001$. Despite the apparent strong result for H_2, no formal favorable statistical conclusion can be made for this second hypothesis, because H_1 is not rejected at conventional significance levels like $\alpha = 0.05$. If this happens in a clinical trial, then non-statisticians would dispute this statistical recommendation; see Fisher and Moyé (1999) for related discussions.

The fallback method was proposed to address this drawback of the fixed-sequence test strategy. In this method, one assigns a portion $\alpha_1 \leq \alpha$ to the first hypothesis in the test sequence, H_1, and distributes the remaining significance level $\alpha - \alpha_1$ to the remaining hypotheses. For example, in testing three hypotheses with the testing sequence $H_1 \rightarrow H_2 \rightarrow H_3$ at $\alpha = 0.05$, one may assign $\alpha_1 = 0.03$ to H_1, $\alpha_2 = 0.01$ to H_2, and $\alpha_3 = 0.01$ to H_3, so that $\alpha_1 + \alpha_2 + \alpha_3 = \alpha$ (see Fig. 6.2). In using this method, the rejection of the hypothesis H_j in the sequence preserves its significance level d_j as the testing progresses and recycles it to the next hypothesis H_{j+1} in the testing sequence, as in the case for the fixed sequential testing method. This recycled significance level adds to the prospectively assigned significance level α_{j+1} of H_{j+1} so that $\alpha_j + \alpha_{j+1}$ becomes the updated significance level for testing H_{j+1}. However, if the procedure fails to reject H_j, then the original significance level α_{j+1} for H_{j+1} remains as is.

Thus, in the above example with three hypotheses, with the testing sequence as stated above, the fallback method would test H_1 at level $\alpha_1 = 0.03$. If this hypothesis is rejected, then the method saves this significance level and recycles it to test H_2 at level $\alpha_1 + \alpha_2 = 0.03 + 0.01 = 0.04$. If H_2 is now rejected at this level, then H_3 is tested at the full significance level $\alpha = 0.05$. However, if H_2 is not rejected at level 0.04, then H_3 is still tested at its original level $\alpha_3 = 0.01$. Consider now the special case that the fallback method is unable to reject H_1 and H_2, but is able to reject H_3. Then, $\alpha_3 = 0.01$ can be recycled to H_1 and H_2, so that these two hypotheses can be tested at higher significance levels. For example, H_1 can be tested at level $\alpha_1 + r\alpha_3 = 0.03 + r\ 0.01$ and H_2 at level $\alpha_2 + (1 - r)\alpha_3 = 0.01 + (1 - r)\ 0.01$, where $0 \leq r \leq 1$ is prospectively specified (see Fig. 6.3a). The resulting test procedure with $r = \alpha_2/(\alpha_1 + \alpha_2)$ is equivalent to the α-exhaustive extension of the fallback procedure considered by Wiens and Dmitrienko (2005). However, Hommel et al. (2007) suggested another extension of the fallback procedure. Figure 6.3b, given in Bretz et al. (2009), displays this extension. In this figure, the symbol ϵ denotes an infinitesimally small weight indicating that the significance level is to be recycled from H_2 to H_3 only when both H_1 and H_2 are rejected. This extension considers that

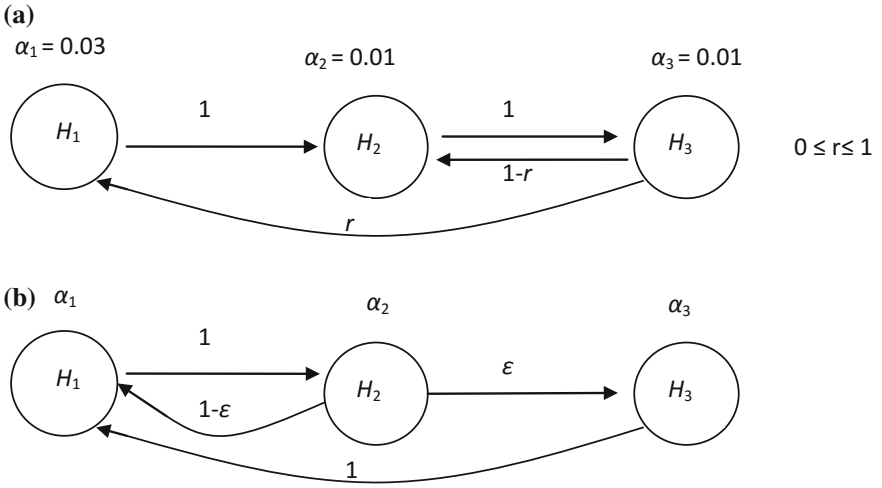


Fig. 6.3 **a**, **b** Graphical illustrations of the two extensions of the fallback procedure; $\alpha_1 + \alpha_2 + \alpha_3 = \alpha = 0.05$

H_1 is more important than H_3; consequently, once H_2 is rejected, its significance level should be recycled to H_1 before continuing to test H_3.

6.2.2 *Recycling of α in Truncated Holm and Hochberg Procedures*

There is a keen interest in using the Holm (1979) and Hochberg procedures (1988) for testing a family of primary hypotheses of clinical trials as these methods are more powerful than the Bonferroni procedure. In order to explain these two procedures, let p_j be the p-value associated with test of hypothesis H_j for $j = 1, \ldots, k$. Further, let $p_{(1)} \leq \ldots \leq p_{(k)}$ be the ordered p-values and $H_{(1)}, \ldots, H_{(k)}$ the associated hypotheses. Assume that these hypotheses are equally weighted. Both procedures use the same test critical values except that the Holm procedure is a step-down procedure and starts testing with the most significant p-value; the Hochberg procedure, on the other hand, is a step-up procedure and starts testing with the least significant p-value.

The Holm procedure follows the following stepwise algorithm:

Step 1: If $p_{(1)} < \alpha/k$, then reject $H_{(1)}$ and go to the next step; otherwise, retain all hypotheses and stop testing.
Step 2: $j = 2, \ldots, k-1$: If $p_{(j)} < \alpha/(k-j+1)$, then reject $H_{(j)}$ and go to the next step; otherwise, retain $H_{(j)}, \ldots, H_{(k)}$ and stop testing.
Step 3: If $k : p_{(k)} < \alpha$, then reject $H_{(k)}$; otherwise, retain it.

The Hochberg procedure follows the following stepwise algorithm:

Step 1: If $p_{(k)} \geq \alpha$, then retain $H_{(k)}$ and go the next step; otherwise, reject all hypotheses and stop testing.
Step 2: $j = 2, \ldots, k-1$: If $p_{(k-j+1)} \geq \alpha/j$, then retain $H_{(k-j+1)}$ and go to the next step; otherwise, reject all remaining hypotheses and stop testing.
Step 3: If $k : p_{(1)} \geq \alpha/k$, then retain $H_{(1)}$; otherwise, reject it.

However, Fig. 6.1 shows that the Holm procedure recycles its entire significance level α within this family. Thus, it is unable to recycle any part of significance level α to other hypotheses unless H_1 and H_2 are both rejected. Consider a trial with a primary family F_1 and a secondary family F_2, where F_1 consists of the two hypotheses H_1 and H_2 and F_2 consists of a single hypothesis H_3. Suppose that one applies the Holm procedure to H_1 and H_2 in F_1 at level $\alpha = 0.05$. If the procedure rejects both H_1 and H_2, then the entire significance level $\alpha = 0.05$ is available for testing H_3 in F_2. However, if it rejects only one hypothesis in F_1, then H_3 in F_2 cannot be tested further; doing so can inflate the FWER. In order to see this, consider the case when H_1 is false with a very large effect so that the procedure rejects H_1 almost surely at level $\alpha/2$, but H_2 and H_3 are both true. Therefore, if H_3 is tested at some level $\alpha_3 > 0$, then the probability of falsely rejecting either H_2 or H_3 will be $1-(1-\alpha)(1-\alpha_3) > \alpha$, assuming independent p-values for H_2 and H_3. For example, if $\alpha_3 = \alpha/2 = 0.025$, then the FWER becomes 0.074 which exceeds the nominal level $\alpha = 0.05$. The same situation holds when the Hochberg procedure is used for the primary family.

Dmitrienko et al. (2008) devised truncated versions of the Holm and Hochberg procedures when there is a desire to have the power advantage of the conventional Holm or Hochberg procedures over the Bonferroni test for testing hypotheses in the primary family but also an interest to retain a portion of the overall significance level α for testing hypotheses in the secondary family, when at least one hypothesis in the primary family is rejected. The truncated versions of the Holm and Hochberg procedures are formed by constructing the critical values c_j for $j = 1, \ldots, k$ by testing its k-ordered hypotheses $H_{(1)}, \ldots, H_{(k)}$ as follows:

$$c_j = \delta \frac{\alpha}{k-j+1} + (1-\delta)\frac{\alpha}{k}, \quad j = 1, \ldots, k, \tag{6.2.1}$$

where the term $\alpha/(k-j+1)$ is the jth critical value for the conventional Holm procedure and the term α/k is the critical value for the Bonferroni test. These two critical values are combined by the parameter δ with $0 \leq \delta \leq 1$ forming a convex combination. Thus, when $\delta = 0$, the method reduces to the Bonferroni method, and when $\delta = 1$ it reduces back to the conventional Holm procedure. The method of carrying on the test procedure remains the same as in the conventional procedure but using the critical values c_j's as defined in (6.2.1). The c_j values remain the same for the Hochberg procedure. However, the test in the Holm procedure starts with the smallest ordered p-value as compared to the Hochberg procedure in which the test starts with the largest ordered p-value.

As an example, consider $k = 2$ and $\delta = 1/2$. Suppose that the p-values for H_1 and H_2 in primary family F_1 are ordered as $p_{(1)} \leq p_{(2)}$, with associated hypotheses ordered as $H_{(1)}$ and $H_{(2)}$. The tests for the truncated Holm procedure are then performed

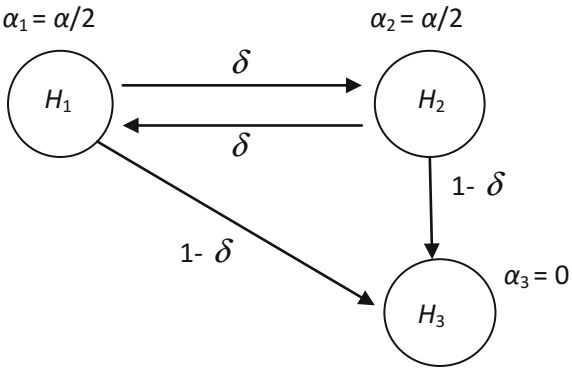


Fig. 6.4 Illustration of the truncated Holm test with two primary hypotheses H_1 and H_2, and one secondary hypothesis H_3, when $\alpha = 0.05$

as follows: (1) Reject $H_{(1)}$ if $p_{(1)} < c_1 = \alpha/2$; otherwise, stop further testing; (2) reject $H_{(2)}$ if $p_{(2)} < c_2 = (1 + \delta)\alpha/2 = 3\,\alpha/4$. The recycled significance level for testing the secondary family F_2 is then either α if both $H_{(1)}$ and $H_{(2)}$ are rejected or $\alpha - c_2 = \alpha/4$ when $H_{(1)}$ is rejected but $H_{(2)}$ is retained.

The truncated Holm procedure can also be graphically represented as in Fig. 6.4. Note that, as compared to Fig. 6.1, the weights on directional edges connecting H_1 with H_2 and vice versa are now δ instead of 1. Choosing $0 \leq \delta < 1$ allows recycling a fraction of the significance level from the primary to the secondary family. Note that if $\delta = 0$ one would recycle $\alpha/2$ to the secondary family, and when $\delta = 1$, then no recycling to the secondary family is foreseen. If the Hochberg procedure is used in this example, then the test procedure would be performed as follows: Reject both $H_{(1)}$ and $H_{(2)}$ if $p_{(2)} < 3\alpha/4$; otherwise, move to test $H_{(1)}$ and reject it if $p_{(1)} < c_1 = \alpha/2$. Huque et al. (2013) discuss truncated Holm procedure for the case of testing three primary hypotheses.

6.3 Closed Testing Procedures

In testing multiple hypotheses in confirmatory clinical trials, the focus generally is on making conclusions for the individual hypotheses; hardly ever one is interested in whether all hypotheses are jointly true or not. Therefore, given k individual hypotheses $H_1, \ldots, H_k$, the test of the global intersection hypothesis $H_{\mathbf{I}} = \bigcap_{j \in \mathbf{I}} H_j$ with $\mathbf{I} = \{1, \ldots, k\}$ is not helpful for this purpose. The rejection of such a global test does not in general help clinicians in characterizing the clinical benefits of the study treatment. However, the *closed testing principle* by Marcus et al. (1976) provides a general framework for constructing powerful test procedures that allow conclusions for the individual hypothesis H_j for $j \in \mathbf{I}$ by testing each intersection hypothesis $H_{\mathbf{J}}$ for $\mathbf{J} \subseteq \mathbf{I}$ at level α. A test procedure based on the closed testing principle is called a closed test procedure and abbreviated by CTP. A CTP for making conclusion on the k individual hypotheses can be constructed as follows:

(1) Define a family $\{H_1, \ldots, H_k\}$ of k individual hypotheses, which can include, for example, primary and secondary hypotheses.
(2) Construct the closure $\tilde{\mathbf{H}}$ of $2^k - 1$ non-empty intersection hypotheses as

$$\tilde{\mathbf{H}} = \left\{ H_{\mathbf{J}} = \bigcap\nolimits_{j \in \mathbf{J}} H_j, \quad \mathbf{J} \subseteq \mathbf{I}, \quad H_{\mathbf{J}} \neq \emptyset \right\}. \tag{6.3.1}$$

(3) Find a suitable α-level test for each intersection hypothesis $H_{\mathbf{J}}$ in $\tilde{\mathbf{H}}$.
(4) Reject an individual hypothesis H_j while controlling the FWER at level α if all $H_{\mathbf{J}}$ in $\tilde{\mathbf{H}}$ with $j \in \mathbf{J}$ are rejected.

The above procedure may sound somewhat complicated, but Table 6.1 illustrates how a CTP can be constructed based on weighted Bonferroni tests for the intersection hypotheses, each tested at level α. This table considers three individual hypotheses H_1, H_2, and H_3, with H_1 and H_2 as primary, and H_3 as secondary. We assume that H_3 is tested only if either H_1 or H_2 is rejected first. Column 1 in Table 6.1 lists all seven intersection hypotheses $H_{\mathbf{J}}$ in $\tilde{\mathbf{H}}$. Columns 2–4 show the pre-specified weights for performing the weighted Bonferroni tests for each intersection hypothesis $H_{\mathbf{J}}$ in Column 1, with the sum of the weights being equal to 1. Column 5 shows the rejection rules of the weighted Bonferroni test for $H_{\mathbf{J}}$ in column 1 based on the p-values p_i for $i = 1, 2, 3$. The CTP from Table 6.1 rejects H_1 while controlling the FWER at level α if it rejects each of the four hypotheses H_{123}, H_{12}, H_{13}, and H_1 locally at level α. Likewise, it rejects H_2 if it rejects H_{123}, H_{12}, H_{23}, and H_2 locally at level α and so on for H_3. Note that if the CTP from Table 6.1 rejects H_1 then it would test only H_{23} and H_2 each at level α to reject H_2, and it would test only H_3 at level α after it rejects both H_1 and H_2.

As mentioned above, the CTP provides strong FWER control in making conclusions about the individual hypotheses. Consider the case of testing, for example, three hypotheses H_1, H_2, and H_3. The null hypothesis configurations for testing these hypotheses are: (1) $H_1H_2H_3$, (2) $H_1H_2K_3$, (3) $H_1K_2H_3$, (4) $K_1H_2H_3$, (5) $H_1K_2K_3$, (6) $K_1H_2K_3$, and (7) $K_1K_2H_3$, where K_j denotes the alternative hypothesis for $j = 1, 2, 3$. For configurations (5)–(7), the procedure obviously controls the Type I error rate at level α, because in each of these cases only one true hypothesis can be falsely rejected and the probability for that is bounded by α. For the configuration $K_1H_2H_3$, there are three ways one can commit a Type I error, namely by falsely rejecting either H_2, H_3 or both H_2 and H_3. Therefore, for this configuration, one must show that the probability $P_{23} = \Pr\{(\text{CTP rejects } H_2) \cup (\text{CTP rejects } H_3) \mid K_1H_2H_3\} \leq \alpha$. Now, (CTP rejects H_2) = $\{(H_{123}$ is rejected) $\cap$ (H_{12} is rejected) $\cap$ (H_{23} is rejected) $\cap$ (H_2 is rejected)$\} \subseteq$ (H_{23} is rejected). Also, (CTP rejects H_3) = $\{(H_{123}$ is rejected) $\cap$ (H_{13} is rejected) $\cap$ (H_{23} is rejected) $\cap$ (H_3 is rejected)$\} \subseteq$ (H_{23} is rejected). Therefore, $P_{23} \leq \Pr\{H_{23} \text{ is rejected}\} \leq \alpha$. Similarly, one can show that for the null hypothesis configurations $H_1H_2K_3$, $H_1K_2H_3$, and $H_1H_2H_3$, the respective probabilities P_{12}, P_{23}, and P_{123} are each bounded by α. Thus, the procedure provides strong FWER controls for the case considered. This proof easily extends to the general case.

Table 6.1 An example of closed testing for three hypotheses with weighted Bonferroni tests

$H_{\mathbf{J}}$	H_1	H_2	H_3	Reject $H_{\mathbf{J}}$ if
H_{123}	0.8	0.2	0.0	$p_j < w_{j;\mathbf{J}}\alpha$ for at least one $j \in \mathbf{J} = \{1,\ 2,\ 3\}$
H_{12}	0.8	0.2	–	$p_j < w_{j;\mathbf{J}}\alpha$ for at least one $j \in \mathbf{J} = \{1,\ 2\}$
H_{13}	0.7	–	0.3	$p_j < w_{j;\mathbf{J}}\alpha$ for at least one $j \in \mathbf{J} = \{1,\ 3\}$
H_1	1.0	–	–	$p_1 < \alpha$
H_{23}	–	0.3	0.7	$p_j < w_{j;\mathbf{J}}\alpha$ for at least one $j \in \mathbf{J} = \{2,\ 3\}$
H_2	–	1.0	–	$p_2 < \alpha$
H_3	–	–	1.0	$p_3 < \alpha$

Note For each $H_{\mathbf{J}}$ in Column 1, the weights $w_{j;\mathbf{J}}$ are such that $0 \leq w_{j;\mathbf{J}} \leq 1$ and the sum $\sum_{j\in\mathbf{J}} w_{j;\mathbf{J}} \leq 1$

6.4 Shortcut Closed Testing Procedures

Although CTPs constructed as above provide powerful multiple test procedures for testing k individual hypotheses, they have the disadvantage that the number of intersection hypotheses increases exponentially in k. Instead, shortcuts to CTPs can be derived that reduce the numerical complexity by exploiting the *consonance* property introduced by Gabriel et al. (1969). A CTP is called consonant if the rejection of an intersection hypothesis implies the rejection of at least one of its individual hypotheses. That is, the rejection of an intersection hypothesis $H_{\mathbf{J}} = \bigcap_{j\in\mathbf{J}} H_j$ at level α implies the rejection of the individual hypothesis H_j at level α for at least one $j \in \mathbf{J}$ while controlling the FWER in the strong sense. The following discusses the general case for constructing shortcut procedures and then the special case using weighted Bonferroni tests. In both cases, one can see that after the rejection of one or more hypotheses using a shortcut procedure the remaining hypotheses can be tested at higher significance levels, though not exceeding α.

6.4.1 The General Case

In order to explain the construction of a shortcut CTP, consider first the case of a conventional CTP for testing three hypotheses H_j with unadjusted p-values p_j for

$j = 1, 2, 3$. One can construct a test for H_{123} by finding significance levels $\alpha_{j;123}$ so that

$$\Pr\left\{\bigcup_{j\in\{1,2,3\}}\left(P_j < \alpha_{j;123}\right)\right\} \leq \alpha. \tag{6.4.1}$$

One would then reject H_{123} at level α when $p_j < \alpha_{j;123}$ for at least one $j \in \{1, 2, 3\}$. Similarly, one can construct tests for H_{12}, H_{13}, and H_{23} by finding significance levels $\alpha_{j;12}$, $\alpha_{j;13}$, and $\alpha_{j;23}$ so that

$$\begin{aligned}&\Pr\left\{\bigcup_{j\in\{1,2,\}}\left(P_j < \alpha_{j;12}\right)\right\} \leq \alpha,\ \Pr\left\{\bigcup_{j\in\{1,3\}}\left(P_j < \alpha_{\alpha_{j;13}}\right)\right\} \leq \alpha, \text{ and}\\ &\Pr\left\{\bigcup_{j\in\{2,3\}}\left(P_j < \alpha_{j;23}\right)\right\} \leq \alpha, \text{ respectively.}\end{aligned} \tag{6.4.2}$$

One would then reject H_{12} when $p_j < \alpha_{j;12}$ for at least one $j \in \{1, 2\}$. Similarly, one would reject H_{13} when $p_j < \alpha_{j;13}$ for at least one $j \in \{1, 3\}$ and would reject H_{23} when $p_j < \alpha_{j;23}$ for at least one $j \in \{2, 3\}$.

Therefore, with the above significance levels for testing intersection hypotheses, one can set up a conventional CTP for testing H_j for $j = 1, 2, 3$. This CTP would reject H_2 if H_{123}, H_{12}, and H_{23} are rejected as above and $p_2 < \alpha$. Similarly, it would reject H_1 if H_{123}, H_{12}, and H_{13} are rejected as above and $p_1 < \alpha$, and so on. However, this CTP simplifies to a shortcut CTP if significance levels in Eq. (6.4.2) also satisfy the conditions:

$$\begin{aligned}&\alpha_{j;123} \leq \alpha_{j;12} \leq \alpha \text{ for } j \in \{1, 2\},\ \alpha_{j;123} \leq \alpha_{j;13} \leq \alpha \text{ for } j \in \{1, 3\}, \text{ and}\\ &\alpha_{j;123} \leq \alpha_{j;23} \leq \alpha \text{ for } j \in \{2, 3\}.\end{aligned} \tag{6.4.3}$$

Conditions (6.4.3) ensure consonance; that is, if an intersection hypothesis is rejected, then at least one of its individual hypotheses is also rejected. Thus, in our case, if H_{123} is rejected by observing $p_2 < \alpha_{2;123}$, then H_2 is rejected—no need to test for H_{12}, H_{23}, and H_2 as in the convention CTP. Now, suppose that H_2 is rejected, and then one would move to test H_{13} using significance levels $\alpha_{j;13}$ for $j \in \{1, 3\}$. If H_{13} is now rejected by observing $p_3 < \alpha_{3;13}$, then H_3 would be rejected. If H_{13} is so rejected, then H_2 would be tested at the full significance level α.

Thus, a CTP would simplify to a shortcut test procedure test if one constructs tests for its intersection hypotheses $H_{\mathbf{J}} = \bigcup_{j\in\mathbf{J}} H_j$ such that these tests satisfy consonance. This can be achieved, for example, by finding significance levels $\alpha_{j;\mathbf{J}}$ such that $\Pr\left\{\bigcup_{j\in\mathbf{J}}\left(P_j < \alpha_{j;\mathbf{J}}\right)\right\} \leq \alpha$ and $\alpha_{j;\mathbf{J}} \leq \alpha_{j;\mathbf{J}'} \leq \alpha$ for all $\mathbf{J}' \subseteq \mathbf{J} \subseteq \mathbf{I}$ and $j \in \mathbf{J}'$. Then, $H_{\mathbf{J}'}$ would be rejected for any $\mathbf{J}' \subseteq \mathbf{J} \subseteq \mathbf{I}$ when $p_j < \alpha_{j;\mathbf{J}'}$ for at least one $j \in \mathbf{J}'$, where, as before, p_j is the unadjusted p-value associated with H_j. Once $H_{\mathbf{J}'}$ is rejected, say for $j = j_0$, then the individual hypotheses H_{j_0} would be rejected, without going through the additional steps of the conventional CTP. We will revisit the special case again in the next section when applying weighted

Bonferroni tests. Note that the above conditions imply that once the procedure rejects an individual hypothesis, the tests for subsequent individual hypotheses can occur at larger significance levels. Satisfying consonance in a CTP has the advantage of reducing the number of tests and simplifying the presentation of the results. The total number of tests in shortcut procedures satisfying consonance reduces the initial complexity from $2^k - 1$ to k.

Example 1 Consider a four-arm trial to compare a treatment with three doses with placebo for showing benefit of the treatment for at least one of the three doses. Let these three doses be denoted as D_1 (high dose), D_2 (medium dose), and D_3 (low dose). Let H_1, H_2, and H_3 denote the corresponding superiority hypotheses by one-sided tests using statistics $Z_j = d_j\sqrt{I_j}$ for $j = 1, 2, 3$. Here, d_j denotes the treatment difference for dose D_j from placebo, $I_j = 1/\text{Var}\ (d_j)$, and $Z_j > 0$ shows that the observed treatment difference is in the beneficial direction. Furthermore, let the sample size n per treatment arm be sufficiently large so that under the null hypothesis H_{123} the joint distribution of (Z_1, Z_2, Z_3) can be assumed to follow a three-dimensional multivariate normal $N_3(\mathbf{0},\ \mathbf{R})$ with mean vector $\mathbf{0}$ and the equi-correlation matrix $\mathbf{R}$ with off-diagonal elements equal to 0.5.

Suppose that in testing H_{123} the significance level $\alpha_{1;\,123}$ is prefixed at 0.6α. The values $\alpha_{2;\,123} = \alpha_{3;\,123} = \alpha_{123}$ have to be determined as follows. Let $c_{j;\,123} = \Phi^{-1}\left(1 - \alpha_{j;\,123}\right)$ for $j = 1, 2, 3$. Then, α_{123} can be obtained from the equation

$$1 - \alpha = \Pr\left\{(Z_1 \le c_{1;\,123}) \cap (Z_2 \le x) \cap (Z_3 \le x) | H_{123}\right\}, \tag{6.4.4}$$

on solving for x, where $\alpha_{123} = 1 - \Phi(x)$ and $\Phi(u)$ denote the distribution function of a standard normal random variable U. Because $\mathbf{R}$ is equi-correlated, consider the test statistics (Z_1, Z_2, Z_3, Y) such that they follow a four-dimensional multivariate normal distribution such that correlations between Z_j and Y are $\sqrt{\rho}$ for $j = 1, 2, 3$. Given $Y = y$, the test statistics Z_j for $j = 1, 2, 3$ are then independently distributed with means $E(Z_j|y) = y\sqrt{\rho}$ and $Var(Z_j|y) = 1 - \rho$. Therefore, x in Eq. (6.4.4) can be determined from the equation

$$1 - \alpha = \int_{-\infty}^{\infty} \left\{\Phi\left(\frac{c_{1;\,123} - y\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right\}\left\{\Phi\left(\frac{x - y\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right\}^2 \varphi(y)dy, \tag{6.4.5}$$

where $\varphi(u)$ is the probability density function of the standard normal random variable U. Equation (6.4.5) involves the evaluation of a one-dimensional integral, which can be computed with, for example, the QUAD function in SAS. Therefore, when $\alpha = 0.025$, one computes $x = 2.48635$ such that $\alpha_{2;\,123} = \alpha_{3;\,123} = a_{123} = 1 - \Phi(x) = 0.006453$. Then, $\alpha_{1;\,123} = 0.015$ and $\alpha_{2;\,123} = \alpha_{3;\,123} = 0.006453$ for testing H_{123}.

Suppose that unadjusted one-sided p-values $p_1 = 0.010$, $p_2 = 0.013$, and $p_3 = 0.034$ were observed in the trial for D_1, D_2, and D_3, respectively. The procedure would then reject H_{123} at level 0.025 as well as the elementary hypothesis H_1 because $p_1 = 0.010 < 0.6\,\alpha = 0.015$; this is possible when one intends to keep the subsequent

significance levels $\alpha_{1;12} = \alpha_{1;13} = \alpha_{1;123}$. Therefore, the procedure would continue testing the intersection hypothesis H_{23} with $\alpha_{2;23} = \alpha_{3;23}$ which can be determined by solving the equation

$$1 - \alpha = \Pr\{(Z_2 \leq y) \cap (Z_3 \leq y) | H_{23}\} = \Phi_{23}(y, y; \rho) \tag{6.4.6}$$

for y where, $\Phi_{23}(u, v; \rho)$ denotes the cumulative distribution function for the standard bivariate normal distribution in the random variables U and V. Solving Eq. (6.4.6) for $\alpha = 0.025$ and $\rho = 0.5$ gives $y = 2.21214$ and $\alpha_{2;23} = \alpha_{3;23} = 1 - \Phi(y) = 0.013479$. Therefore, the procedure would also reject H_{23} at level $\alpha = 0.025$ (one-sided) and with its individual hypothesis H_2, because $p_2 = 0.013 < 0.013479$. The procedure would then continue testing H_3 and would fail to reject it because $p_3 = 0.034 > 0.025$.

6.4.2 *Special Case Using Bonferroni Tests*

An important class of closed test procedures is derived by applying the weighted Bonferroni test to each intersection hypothesis $H_{\mathbf{J}} = \bigcap_{j \in \mathbf{J}} H_j$ with $\mathbf{J} \subseteq \mathbf{I}$. Accordingly, one specifies the weights $w_{j;\mathbf{J}}$ such that $0 \leq w_{j;\mathbf{J}} \leq 1$ and $\sum_{j \in \mathbf{J}} w_{j;\mathbf{J}} \leq 1$. Using the weighed Bonferroni test, one rejects $H_{\mathbf{J}}$ if $p_j < \alpha_{j;\mathbf{J}} = w_{j;\mathbf{J}}\alpha$ for at least one $j \in \mathbf{J}$. Hommel et al. (2007) introduced a subclass of sequentially rejective closed test procedures by imposing the monotonicity condition

$$w_{j;\mathbf{J}} \leq w_{j;\mathbf{J}'} \leq 1 \text{ for all } \mathbf{J}' \subseteq \mathbf{J} \subseteq \mathbf{I}, \text{ and } j \in \mathbf{J}' \tag{6.4.7}$$

on the weights of the weighted Bonferroni tests. They showed that by testing intersection hypotheses with weighted Bonferroni tests satisfying (6.4.7) guarantees consonance. That is, the weighted Bonferroni test rejects an intersection hypothesis $H_{\mathbf{J}}$ if $p_j < \alpha_{j;\mathbf{J}} = w_{j;\mathbf{J}}\alpha$ for at least one $j \in \mathbf{J}$ and consequentially also the elementary hypothesis H_j if the weights satisfy (6.4.7). This leads to a simplified shortcut procure to the underlying CTP. Many commonly used multiple test procedures satisfy condition (6.4.7), including the weighted Bonferroni–Holm procedure, certain gatekeeping procedures, fixed-sequence tests, the fallback procedure, and also the graphical approach by Bretz et al. (2009).

However, conventional CTPs based on the weighted Bonferroni weights will still be valid even if the weights do not satisfy the consonance condition. For example, in Table 6.1, the Bonferroni weights do not satisfy condition (6.4.7). The weights $\{w_{j;\mathbf{J}}\}$ for $\mathbf{J} = \{1, 2, 3\}$ are $\{0.8, 0.2, \text{and } 0.0\}$ and the weights $\{w_{j;\mathbf{J}'}\}$ for $\mathbf{J}' = \{1, 2, 3\}$ are $\{0.7 \text{ and } 0.3\}$, violating the condition (6.4.7), because $w_{1;\mathbf{J}'} = 0.7 < 0.8 = w_{1;\mathbf{J}}$. This type of weighting or others which do not satisfy consonance or satisfy only partially is beyond the scope of this chapter. However, the weights in Table 6.1 can be easily modified to satisfy (6.4.7) by setting, for example, $w_{1;\mathbf{J}'} = 0.8$ and $w_{3;\mathbf{J}'} = 0.2$. The following illustrates how to select weights for the weighted Bonferroni tests that satisfy (6.4.7) and lead to shortcut CTPs for clinical trials applications.

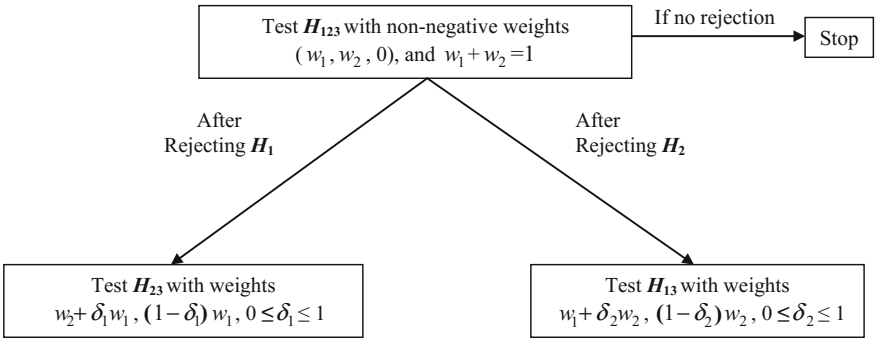


Fig. 6.5 A shortcut CTP representation for testing two primary hypotheses H_1 and H_2, and a secondary hypothesis H_3 on using the weighted Bonferroni tests for intersection hypotheses with weights satisfying the consonance condition. Values of the recycling parameters δ_1 and δ_2 in this figure are pre-specified

Consider first the simple case of testing three hypotheses, where the first two hypotheses H_1 and H_2 are of primary and the third one, H_3, is of secondary interest. Suppose that in testing the initial intersection hypothesis H_{123} one selects the weights of w_1, w_2, and w_3 for H_1,H_2, and H_3, respectively, such that $w_1 + w_2 = 1$ and $w_3 = 0$. The selection of $w_3 = 0$ for H_3 indicates that this hypothesis is tested only after rejecting at least one of the two primary hypotheses.

Suppose that $p_1 < w_1\alpha$ and the resulting weighted Bonferroni test rejects H_{123}. This rejection then implies the rejection of the individual hypothesis H_1 such that its weight w_1 can be recycled to the remaining weights w_2 and w_3 according to a parameter $\delta_1 \geq 0$, so that in testing the intersection hypothesis H_{23} the updated weights become $w_2 + \delta_1 w_1$ and $w_3 + (1 - \delta_1) w_1$. Note that in this strategy for updating weights, none of the weights in testing H_{23} are smaller than the corresponding weights in testing H_{123}. Similar arguments apply when the initial test for H_{123} rejects H_2 instead of H_1. In that case, the updated weights for testing H_{13} become $w_1 + \delta_2 w_2$ and $w_3 + (1 - \delta_2) w_2$ according to a recycling parameter $\delta_2 \geq 0$. Consequently, the weighting structure for the intersection hypotheses is such that consonance is ensured. Therefore, the shortcut procedure as displayed in Fig. 6.5 can be created for testing H_1, H_2, H_3.

Now, consider a trial to test two primary hypotheses H_1 and H_2, and two secondary hypotheses H_3 and H_4 using weighted Bonferroni tests, where H_3 and H_4 are paired with H_1 and H_2, respectively, so that H_3 can be tested only after H_1 is first tested and rejected; similarly,H_4 can be tested only after H_2 is first tested and rejected. Following Maurer et al. (2011), we will call $\{H_1, H_3\}$ and $\{H_2, H_4\}$ as pairs of parent–descendant hypotheses to reflect the above-stated hierarchical testing strategy. Suppose that H_1 and H_2 receive the initial weights w_1 and w_2, respectively, such that $w_1 + w_2 = 1$, so that $w_3 = w_4 = 0$ for H_3 and H_4, respectively. The selections $w_3 = 0$ and $w_4 = 0$ indicate that a descendant secondary hypothesis is not to be tested until its parent primary hypothesis is first tested and rejected. These weights represent the

Table 6.2 Bonferroni weights with recycling parameters δ_1 and δ_2 for testing the two primary hypotheses H_1 and H_2, and two secondary hypotheses H_3 and H_4 in performing a shortcut CTP

Intersection hypotheses	Weights assigned to hypotheses for performing weighted Bonferroni tests			
	H_1	H_2	H_3	H_4
H_{1234}	w_1	$w_2 = 1 - w_1$	0	0
H_{234}	–	$w_2 + \delta_1 w_1$	$(1 - \delta_1)w_1$	0
H_{134}	$w_1 + \delta_2 w_2$	–	0	$(1 - \delta_2)w_2$
H_{13}	1	–	0	–
H_{24}	–	1	–	0
H_{34}	–	–	w_3'	w_4'

$w_3' = (w_1 + \delta_2 w_2)/(1 + \delta_1 w_1 + \delta_2 w_2)$ and $w_4' = (w_2 + \delta_1 w_1)/(1 + \delta_1 w_1 + \delta_2 w_2)$

local significance levels $\alpha_j = w_j \alpha$ for $j = 1, \ldots, 4$. As before, consider the recycling parameters δ_1 and δ_2 on the interval [0, 1].

With the above information, one can construct the weights in Table 6.2 for performing weighted Bonferroni tests of intersection hypotheses by considering only the six intersection hypotheses (of order ≥ 2 as listed in Table 6.2) satisfying the consonance condition. Note that in this table the number six results from the weighting strategy if a primary hypothesis is rejected such that a fraction of its weight is propagated to the other primary hypothesis and the remaining weight goes to its descendant secondary hypothesis; see the rows for H_{234} and H_{134} in Table 6.2. For example, if H_1 was initially rejected in testing H_{1234} according to the weights in Row 3 of Table 6.2, then in testing H_{234} in Row 4 of Table 6.2 a fraction δ_1 of w_1 goes to H_2 making the total weight at H_2 as $w_2 + \delta_1 w_1$ and remaining weight $(1 - \delta_1)w_1$ is assigned to H_3 which is descendant of H_1, so that the weight at H_4 still remains 0. Similarly, if H_2 was initially rejected, then in testing H_{134} a fraction δ_2 of w_2 goes to H_1 making the total weight at H_1 as $w_1 + \delta_2 w_2$ and remaining weight $(1 - \delta_2)w_2$ is assigned to H_4, so that the weight at H_3 still remains 0.

For the above weighting strategy, Table 6.2 is sufficient to test the four individual hypotheses, as it spans the full closure satisfying (6.4.7) with the following conditions: (1) The weights for indices in H_{123}, H_{124}, and H_{12} remain the same as the corresponding weights for indices in H_{1234}; (2) the weights for indices in H_{23} remain the same as the corresponding weights for indices in H_{234}; (3) the weights for indices in H_{14} remain the same as the corresponding weights for indices H_{134}. Table 6.3 shows the full closure satisfying (6.4.7) without the four individual hypotheses. Thus, Table 6.2 leads to a shortcut procedure as displayed in Fig. 6.6, where the maximum number of tests is $k = 4$ as compared to the $2^k - 1 = 15$ for the full closure.

The weights shown for H_{34} in Table 6.2 are proportional to $w_1 + \delta_2 w_2$ for the index 3 of the set $\{3, 4\}$ and to $w_2 + \delta_1 w_1$ for the index 4 of this set. Thus, the actual weights for testing H_{34} become $w_3' = (w_1 + \delta_2 w_2)/(1 + \delta_1 w_1 + \delta_2 w_2)$ and

Table 6.3 Bonferroni weights with recycling parameters δ_1 and δ_2 for testing the two primary hypotheses H_1 and H_2, and two secondary hypotheses H_3 and H_4, in the full closure table satisfying consonance (showing only intersection hypotheses of order 2 or greater)

Intersection hypotheses	Weights assigned to hypotheses for performing weighted Bonferroni tests			
	H_1	H_2	H_3	H_4
$\boldsymbol{H}_{1234}$	w_1	$w_2 = 1 - w_1$	0	0
H_{123}	w_1	w_2	0	–
H_{124}	w_1	w_2	–	0
H_{12}	w_1	w_2	–	–
$\boldsymbol{H}_{134}$	$w_1 + \delta_2 w_2$	–	0	$(1 - \delta_2)w_2$
H_{14}	$w_1 + \delta_2 w_2$	–	–	$(1 - \delta_2)w_2$
$\boldsymbol{H}_{13}$	1	–	0	–
$\boldsymbol{H}_{234}$	–	$w_2 + \delta_1 w_1$	$(1 - \delta_1)w_1$	0
H_{23}	–	$w_2 + \delta_1 w_1$	$(1 - \delta_1)w_1$	–
$\boldsymbol{H}_{24}$	–	1	–	0
$\boldsymbol{H}_{34}$	–	–	w_3'	w_4'

Note Intersection hypotheses shown in bold texts are the ones included in Table 6.2

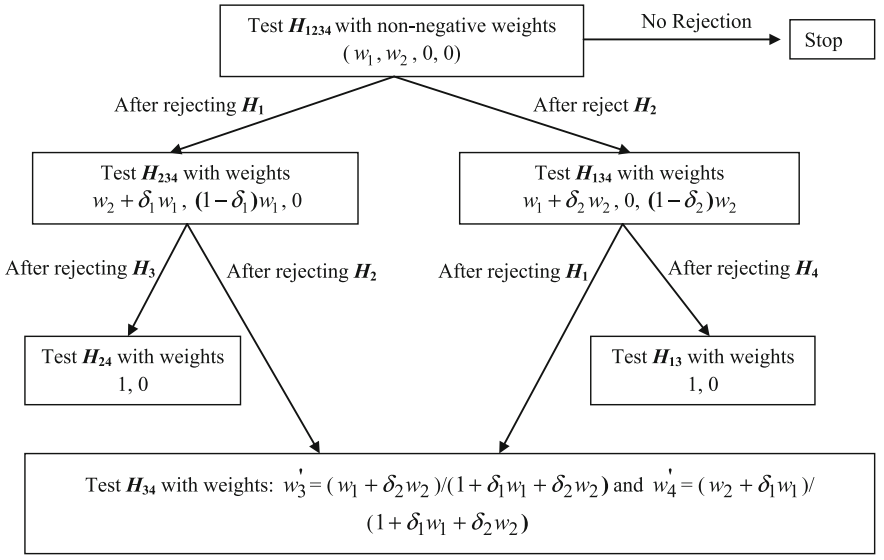


Fig. 6.6 A shortcut CTP representation for testing two primary hypotheses H_1 and H_2, and two secondary hypotheses H_3 and H_4 on using the weighted Bonferroni tests for intersection hypotheses with weights satisfying the consonance condition. Values of δ_1 and δ_2 in this figure are pre-specified

$w_4' = (w_2 + \delta_1 w_1)/(1 + \delta_1 w_1 + \delta_2 w_2)$, respectively, and sum to 1. When $\delta_1 = \delta_2 = \delta$,

Table 6.4 A different table of Bonferroni weights for testing the two primary hypotheses H_1 and H_2, and two secondary hypotheses H_3 and H_4. After rejection of a primary hypothesis, a fraction of its weight goes to the second primary hypothesis and the remaining is distributed equally to the two secondary hypotheses

Intersection hypotheses	Weights assigned to hypotheses for performing weighted Bonferroni tests			
	H_1	H_2	H_3	H_4
H_{1234}	w_1	$w_2 = 1 - w_1$	0	0
H_{134}	$w_1 + \delta_2 w_2$	–	$(1 - \delta_2)w_2/2$	$(1 - \delta_2)w_2/2$
H_{13}	$w_1 + \delta_2 w_2 + (1 - \delta_2)w_2/2$	–	$(1 - \delta_2)w_2/2$	–
H_{14}	$w_1 + \delta_2 w_2 + (1 - \delta_2)w_2/2$	–	–	$(1 - \delta_2)w_2/2$
H_{234}	–	$w_2 + \delta_1 w_1$	$(1 - \delta_1)w_1/2$	$(1 - \delta_1)w_1/2$
H_{23}	–	$w_2 + \delta_1 w_1 + (1 - \delta_1)w_1/2$	$(1 - \delta_1)w_1/2$	–
H_{24}	–	$w_2 + \delta_1 w_1 + (1 - \delta_1)w_1/2$	–	$(1 - \delta_1)w_1/2$
H_{34}	–	–	1/2	1/2

then $w_3' = (w_1 + \delta w_2)/(1 + \delta)$ and $w_4' = (w_2 + \delta w_1)/(1 + \delta)$. Note that consonance is also satisfied if one selects $w_3' = w_1$ and $w_4' = w_2$.

Note that if the weighting strategy was different than in Table 6.2, then one may have to consider more than six intersection hypotheses in order to satisfy (6.4.7). For example, when the pairs $\{H_1, H_3\}$ and $\{H_2, H_4\}$ are not parent–descendant, then one may consider a weighting strategy such that after a primary hypothesis is initially rejected then a fraction of its weight goes to the other primary hypothesis, but the remaining fraction is distributed equally to the two secondary hypotheses. This weighting strategy then generates Table 6.4 of weights for eight intersection hypotheses and the associated test scheme as shown in Fig. 6.7. Such a weighting strategy is followed in the parallel gatekeeping method. Similar to Tables 6.2 and 6.4, span the full closure table satisfying (6.4.7) with the condition that the weights for indices in H_{123}, H_{124}, and H_{12} remain the same as the corresponding weights for indices in H_{1234}.

6.5 Sequentially Rejective (SR) Graphical Procedures

The SR graphical procedures as proposed in Bretz et al. (2009) visualize the Bonferroni-based tests for each individual hypothesis along with an α-propagation rule by which the procedure recycles the significance level of a rejected hypothesis to other remaining hypotheses to be tested. In a graphical approach, the k individual

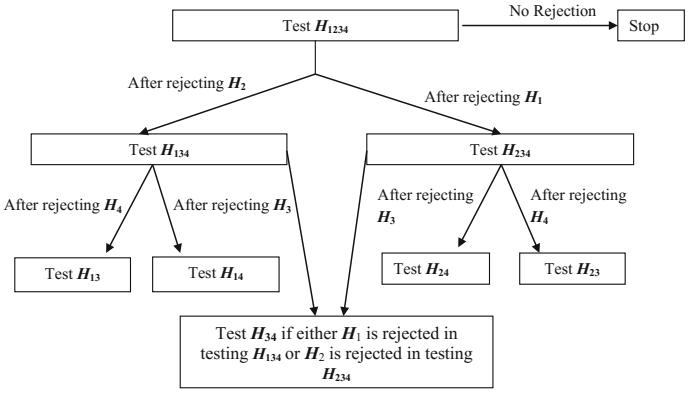


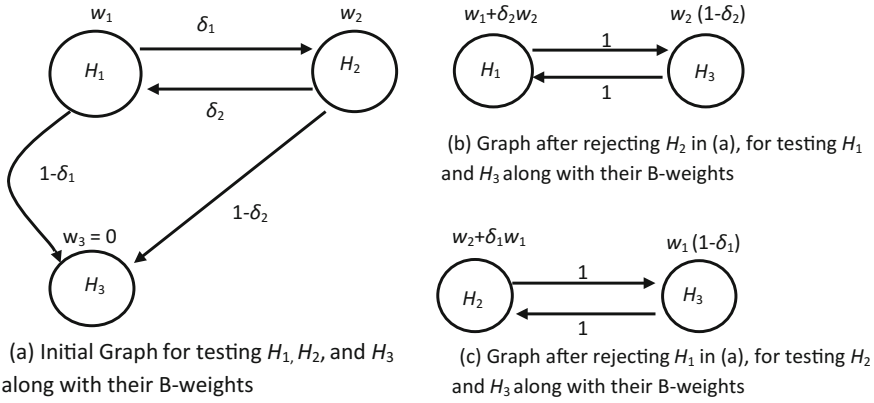
Fig. 6.7 A shortcut CTP representation for testing two primary hypotheses H_1 and H_2, and two secondary hypotheses H_3 and H_4 on using the weighted Bonferroni tests for intersection hypotheses with weights as shown in Table 6.4. Values of δ_1 and δ_2 in this figure are pre-specified

hypotheses are represented initially by a set of k nodes with a nonnegative weight of w_i at node i ($i = 1, \ldots, k$); this weight when multiplied by α represents the local significance level at that node. The weight g_{ij} associated with a directed edge-connecting node i with node j indicates the fraction of the local significance level at the tail node i that is added to the significance level at the terminal node j if the hypothesis at the tail node i is rejected. For convenience, we will call these directed edges as "arrows" running from one node to the other and the weight g_{ij} as the "transition weight" on the arrow running from node i to node j.

Figure 6.8 illustrates the basic concepts for testing two primary hypotheses H_1 and H_2, and one secondary hypothesis H_3. In this figure, the initial Graph (a) shows three nodes. Two nodes represent H_1 and H_2 with weights w_1 and w_2, respectively, with $w_1 + w_2 = 1$. The node for H_3 shows a weight $w_3 = 0$, which can increase only after the rejection of a primary hypothesis. The nonnegative number $g_{12} = \delta_1$ is the transition weight on the arrow going from H_1 to H_2; similarly, $g_{21} = \delta_2$ is the transition weight on the arrow going from H_2 to H_1. The transition weight on the arrow going from H_1 to H_3 is $1 - \delta_1$ and that on the arrow going from H_2 to H_3 is $1 - \delta_2$ satisfying the condition that sum of the transition weights on all outgoing arrows from a single node must be bounded above by 1.

Graph (b) of Fig. 6.8 represents the resulting graph after H_2 is rejected in Graph (a). The rejection of this hypothesis frees its weight w_2 which is then recycled to H_1 and H_3 according to an α-propagation rule addressed in the following for the general case. This rule also calculates new transition weights going from one node to the other for the new graph. Graph (c) of Fig. 6.8 similarly shows the resulting graph if H_1 is rejected in Graph (a). The following shows the general SR graphical procedure for testing k individual hypotheses $H_1, \ldots, H_k$ given their individual unadjusted p-values.

(0) Set $\mathbf{J} = \{1, \ldots, k\}$.



(a) Initial Graph for testing H_1, H_2, and H_3 along with their B-weights

(b) Graph after rejecting H_2 in (a), for testing H_1 and H_3 along with their B-weights

(c) Graph after rejecting H_1 in (a), for testing H_2 and H_3 along with their B-weights

Fig. 6.8 Graphical representation of intersection hypotheses in the shortcut closed testing with two primary hypotheses H_1 and H_2, and one secondary hypothesis H_3; the term B-weights used here is for Bonferroni weights

(i) Select a $j \in \mathbf{J}$ such that $p_j < w_{j;\mathbf{J}}\alpha$, and reject H_j; otherwise, stop. The collection of weights $\{w_{j;\mathbf{J}}, j \in \mathbf{J}\}$ are such that $0 \leq w_{j;\mathbf{J}} \leq 1$ and $\sum_{j\varepsilon\mathbf{J}} w_{j;\mathbf{J}} \leq 1$.

(ii) Update the graph:

(a) $\mathbf{J} = \mathbf{J}\backslash(j)$

(b)

$$w_{l;\mathbf{J}} = w_{l;\mathbf{J}} + w_{j;\mathbf{J}} g_{jl}, \quad l \in \mathbf{J}; \ 0, \text{ otherwise} \tag{6.5.1}$$

(c)

$$g_{lk} = \left\{ \frac{g_{lk} + g_{lj} g_{jk}}{1 - g_{lj} g_{jl}}, \ (l,k) \in \mathbf{J}, \ l \neq k, \ g_{lj} g_{jl} < 1 \right\}; \ 0, \text{ otherwise} \tag{6.5.2}$$

(iii) If $|\mathbf{J}| \geq 1$, then go to step (i); otherwise, stop.

After rejecting H_j, the above (6.5.1) for a new graph updates the weight for H_l to a new weight which is the old weight $w_{l;\mathbf{J}}$ plus the weight $w_{j;\mathbf{J}}$ at H_j multiplied by the transition weight g_{jl} on the arrow connecting H_j to H_l. Also, (6.5.2) calculates new transition weights for this new graph for which the numerator $g_{lk} + g_{lj} g_{jk}$ is the transition weight on the arrow connecting H_l to H_k plus the product of the transition weights on arrows going indirectly from H_l to H_k through the rejected hypothesis H_j. The denominator term $g_{lj} g_{jl}$ in (6.5.2) indicates the product of transition weights on arrows connecting H_l to H_j and returning back to H_l. The procedure produces weights $w_{l;\mathbf{J}}$ which satisfy the consonance condition of (6.4.7).

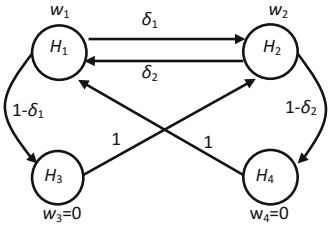
In order to illustrate the graphical procedure, consider an oncology trial for comparing two doses (high and low) of a new treatment to a control on two primary

endpoints progression-free survival (PFS) and overall survival (OS) with the logical restriction that the low dose can be tested for the treatment efficacy on an endpoint only when the high dose for that endpoint first establishes treatment efficacy. Thus, for this trial we have a total of four hypotheses, grouped into two primary hypotheses H_1 and H_2 comparing the high dose to the control for the two endpoints PFS and OS, and the two secondary hypotheses H_3 and H_4 comparing the low dose to the control for the same two endpoints. The logical restriction stated above implies the pairing of hypotheses as $\{H_1, H_3\}$ and $\{H_2, H_4\}$ reflect the hierarchy in testing the two doses for each endpoint; that is, H_3 is to be tested only when H_1 is first tested and rejected, and similarly, H_4 is to be tested when H_2 is first tested and rejected.

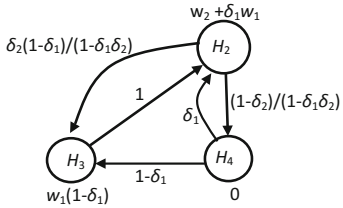
The initial Graph (a) of Fig. 6.9 displays the graphical testing strategy employed for the above trial. The four individual hypotheses H_1, H_2, H_3, and H_4 are represented by four nodes with associated nonnegative weights w_1, w_2, w_3, and w_4, respectively. These weights satisfy $w_1 + w_2 = 1$, $w_3 = 0$, and $w_4 = 0$ so that the local significance levels for testing H_1 and H_2 are $w_1\alpha$ and $w_2\alpha$, respectively, and those of for testing H_3 and H_4 are zero in this graph. The zero weight assignment for the two secondary hypotheses indicates that in this initial graph we do not want to reject a secondary hypothesis until its parent primary hypothesis is first rejected.

In Graph (a) of Fig. 6.9, $g_{12} = \delta_1$, $g_{21} = \delta_2$, $g_{13} = 1 - \delta_1$ and $g_{24} = 1 - \delta_2$. These settings mean that if H_1 was rejected in Graph (a), then a fraction δ_1 of w_1 would recycle H_2 so that the weight at H_2 would become $w_2 + \delta_1 w_1$ and the remainder $(1 - \delta_1)w_1$ would go to H_3; the weight at H_4 would remain 0 because there is no arrow going from H_1 to H_4 setting $g_{14} = 0$. Similarly, if H_2 was initially rejected in Graph (a), then a fraction δ_2 of w_2 would recycle H_1 so that the weight at H_1 would become $w_1 + \delta_2 w_2$ and the remainder $(1 - \delta_2)w_2$ would go to H_4; the weight at H_3 would remain 0 as there is no arrow going from H_2 to H_3 giving $g_{23} = 0$. The value $g_{32} = 1$ indicates that if H_3 was rejected after the rejection of H_1 then the entire weight $(1 - \delta_1)w_1$ at H_3 would recycle H_2, so that the total weight at H_2 after the rejection of both H_1 and H_3 would be $(w_2 + \delta_1 w_1) + (1 - \delta_1)w_1 = 1$. Similarly, $g_{41} = 1$ indicates that if H_4 was rejected after the rejection of H_2 then the entire weight $(1 - \delta_2)w_2$ at H_4 would recycle H_1, so that the total weight at H_1 after the rejection of both H_2 and H_4 would be $(w_1 + \delta_2 w_2) + (1 - \delta_2)w_2 = 1$.

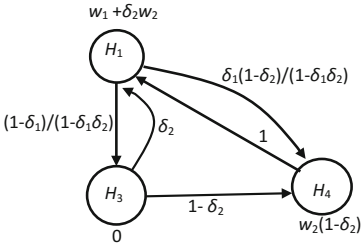
Suppose that in Graph (a) of Fig. 6.9, one selects $w_1 = 4/5$, $w_2 = 1/5$, $\delta_1 = 1/4$, and $\delta_2 = 3/4$. This selection may be based on the prior experience that such a trial can easily succeed for the high dose with the PFS endpoint, and there may be similar experience for the low dose as well. Suppose that H_1 is rejected in this graph, then H_1 is removed, and testing reduces to testing the remaining three hypotheses with a new Graph (b). For this new graph, the application of (6.5.1) yields new weights of $w_2 + \delta_1 w_1 = 2/5$, $(1 - \delta_1)w_1 = 3/5$, and $w_4 = 0$ for H_2, H_3, and H_4, respectively; also, the application of (6.5.2) finds $g_{23} = \delta_2(1 - \delta_1)/(1 - \delta_1\delta_2) = 9/13$, $g_{24} = (1 - \delta_2)/(1 - \delta_1\delta_2) = 4/13$, $g_{32} = 1$, $g_{42} = \delta_1 = 1/4$, and $g_{43} = 1 - \delta_1 = 3/4$. However, if H_2 was rejected instead of H_1 in Graph (a), then one would arrive at Graph (c) with weights $w_1 + \delta_2 w_2 = 19/20$, $w_3 = 0$, and $(1 - \delta_2)w_2 = 1/20$ for H_1, H_3, and H_4, respectively, with $g_{14} = \delta_1(1 - \delta_2)/(1 - \delta_1\delta_2) = 1/13$, $g_{13} = (1 - \delta_1)/(1 - \delta_1\delta_2) = 12/13$, and $g_{41} = 1$, $g_{31} = \delta_2 = 3/4$, and $g_{34} = 1 - \delta_2 = 1/4$.



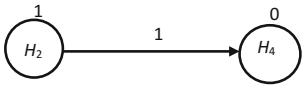
(a) Initial Graph for testing H_1, H_2, H_3, and H_4 along with their weights



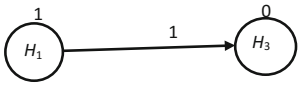
(b) Graph for testing for testing H_2, H_3, and H_4 along with their weights after the rejection of H_1 in (a)



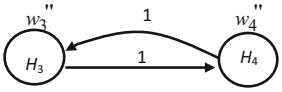
(c) Graph for testing H_1, H_3, and H_4 along with their weights after the rejection of H_2 in (a)



(d) Graph for testing H_2 and H_4 along with their B-weights after the rejection of H_3 in (b)



(e) Graph for testing H_1 and H_3 along with their weights after the rejection of H_4 in (c)



(f) Graph for testing H_3 and H_4 along with their weights after the rejection of H_2 in (b) or H_1 in (c)

$$w_3'' = (w_1 + \delta_2 w_2)\left(\frac{(1-\delta_1)}{1-\delta_1\delta_2}\right)$$

$$w_4'' = (w_2 + \delta_1 w_1)\left(\frac{(1-\delta_2)}{1-\delta_1\delta_2}\right)$$

Fig. 6.9 Graphical representation of the active intersection hypotheses in the shortcut closed testing with two primary hypotheses H_1 and H_2, and two secondary hypotheses H_3 and H_4

Now, in Graph (b), if H_3 is rejected then the application of (6.5.1) and (6.5.2) gives Graph (d) for testing the remaining two hypotheses H_2 and H_4. Similarly, in Graph (c), if H_4 is rejected, then one would arrive at Graph (e) for testing the remaining two hypotheses H_1 and H_3. However, if either in Graph (b) H_2 is rejected or in Graph (c) H_1 is rejected, then one would reach Graph (f) for testing H_3 and H_4, with weights w_3'' and w_4'', respectively, and $g_{34} = g_{43} = 1$, where

$$w_3'' = (w_1 + \delta_2 w_2)\left(\frac{(1-\delta_1)}{1-\delta_1\delta_2}\right) = 57/65 \text{ and } w_4'' = (w_2 + \delta_1 w_1)\left(\frac{(1-\delta_2)}{1-\delta_1\delta_2}\right) = 8/65.$$

The graphical procedure in Fig. 6.9 and that based on the shortcut closed testing as shown in Table 6.2 (and in Fig. 6.6) are two different multiple test procedures for the

same clinical trial problem. However, the graphical procedure shows resemblance to the shortcut closed testing. This can be viewed by considering each graph in the graphical approach with two or more hypotheses as a Bonferroni test for an intersection hypothesis along with an α-propagation rule and a rule that generates weights for the weighted Bonferroni test for the next graph after the rejection of a hypothesis. For example, Graph (a) in Fig. 6.9 displays the weighted Bonferroni test for H_{1234} as well as the recycling rule. Now, if the procedure rejects H_1, then one is Graph (b) of Fig. 6.9, which again represents a weighted Bonferroni test for H_{234} (with the weights that satisfy consonance) along with the recycling rule for the next graph. Thus, on comparing Fig. 6.5 with Fig. 6.8 and Fig. 6.6 with Fig. 6.9, one can see this resemblance.

In addition, to a user it may seem that graphical procedures do not include all shortcut procedures as defined in Sect. 6.4. For example, in the graphical approach for testing H_3 and H_4 (see Graph (f) of Fig. 6.9), the Bonferroni weights associated with H_3 and H_4 on using (6.5.1) and (6.5.2), after rejecting H_1 and H_2, come out to be w_3'' and w_4'', respectively (see bottom of Fig. 6.9). For example, in Fig. 6.9, if $w_1 = 3/4$, $w_2 = 1/4$, $\delta_1 = 3/4$, and $\delta_2 = 1/2$, then for these values in the initial graph $w_3'' = 1/4$ and $w_4'' = 3/4$. Shortcut closed testing, on the other hand (see bottom of Table 6.2), can use these weights as well as other weights, such as $w_3' = 2/5$ and $w_4' = 3/5$ in Table 6.2 for the test of H_{34}, satisfying consonance. But $\delta_1 = \delta_2 = \delta$ gives $w_3'' = w_3'$ and $w_4'' = w_4'$. However, as indicated in Bretz et al. (2011, 2014), the graphical procedure has extensions, thus covering a greater set of shortcut procedures. Also, note that although choosing from a larger class of shortcut procedures would lead to even more possibilities of choosing suitable multiple test procedures, the determination of the weights becomes much more demanding and more difficult to communicate. Thus, in our opinion, the SR graphical approach is flexible and general enough for practical purposes and for clinical trial applications.

6.5.1 Re-testing of a Primary Hypothesis

After the rejection of a primary hypothesis, methods discussed in Sects. 6.2 and 6.4, and the graphical procedures discussed in Sect. 6.5, allow re-testing of other unrejected primary hypotheses of a trial at higher significance levels on recycling the significance level of the rejected primary hypothesis of that trial to these unrejected primary hypotheses according to a pre-specified α-propagation rule. Suppose that the trial also intends to test parent–descendant secondary hypotheses. As defined earlier, a secondary hypothesis of a trial is called a parent–descendant of a primary hypothesis of that trial if that secondary hypothesis can be tested only if its associated primary hypothesis is first tested and rejected. Now, suppose that the investigator after the rejection of a primary hypothesis of the trial recycles the entire significance level of this rejected primary hypothesis to test other primary hypotheses of the trial and fails to reject them. Then in that case, no significance level will be left to test the secondary hypothesis to support the result of the rejected primary hypothesis.

This situation may cause difficulties in interpreting study findings. Therefore, our suggestion is that once a primary hypothesis in a trial is rejected, at least part of its significance level should be recycled to test its associated secondary hypothesis, and if that secondary hypothesis is rejected, then its significance level can be recycled back to test other unrejected primary hypotheses of the trial.

However, in order to increase the power of the test for the secondary hypothesis after the rejection of its associated primary hypothesis, one may recycle the entire significance level of the rejected primary hypothesis to this secondary hypothesis. This seems quite alright, when the test of the associated secondary hypothesis takes priority for enhancing the credibility of the rejected primary hypothesis. Nonetheless, some, being naïve, may refute this approach on stating that the primary hypotheses test results must not depend in any way on the results of the secondary hypotheses test results. However, such logical restriction is apparently unwarranted here as one views the successful outcomes on the test of a primary hypothesis and its associated secondary hypothesis serving as a gate before testing subsequent hypotheses. This seems quite meaningful when testing multiple hypotheses related to different doses or different endpoints, as in the following example.

Example 2 Consider the above three-arm oncology trial with control and two doses D_1 and D_2, where D_1 is of greater potency than D_2, and each dose is compared to the control. As before, let H_1 and H_2 be the two primary hypotheses for demonstrating treatment benefits for the high dose D_1 on endpoints PFS and OS, respectively. Further, let H_3 and H_4 be the two secondary hypotheses for demonstrating treatment benefits for the low dose D_2 on PFS and OS, respectively. The logical restriction imposed is that H_3 is to be tested only after H_1 is first tested and rejected, and similarly, H_4 is to be tested only after H_2 is first tested and rejected. Let the weighted Bonferroni test procedure initially tests H_1 and H_2 at levels 0.04 and 0.01, respectively, and suppose that in this testing, it rejects H_1 but fails to reject H_2. Rejecting H_1 at level 0.04 then allows testing H_3 at this significance level. Now, suppose that H_3 when tested at level 0.04 is rejected. Then, this significance level of 0.04 is free and can be recycled for re-testing H_2 at a much higher significance level of 0.01 + 0.04 = 0.05. Therefore, if H_2 is now rejected, because its test now has greater power, then H_4 can be tested at the full significance level of 0.05. Thus, in this illustration the test result for H_1 is independent of the results of secondary hypotheses. Therefore, if one is able to characterize treatment benefit of the study medication based on this result, then it should be alright to let the result of H_2 depend on the result of H_3 as long as the test procedure applied controls FWER in the strong sense for testing the stated four hypotheses of the trial.

6.6 Discussion

Confirmatory trials generally include multiple objectives, frame these objectives in terms of multiple test hypotheses, and subsequently classify these hypotheses into

hierarchically ordered families. This hierarchical ordering of multiple hypotheses, driven mostly by clinical considerations, plays an important role in the statistical testing, as it reduces the degree of multiplicity and allows the use of methods that are more powerful. The use of powerful statistical tests for primary hypotheses of a trial is of paramount importance, as the trial can become a successful trial if it rejects (i.e., wins for) at least one of its primary hypotheses. On the other hand, a trial that wins for its secondary hypothesis but fails for all its primary hypotheses is usually a failed trial. Therefore, confirmatory trials normally use a statistical test method that assigns initially all of the trial α to the test of the primary hypotheses and in testing gives more weights to those primary hypotheses that are more important than others. Novel statistical test methods (such as gatekeeping and the SR graphical procedures) are now available that do all that.

Statistical tests for intersection hypotheses, often known as global tests, test a global null hypothesis for finding whether given a number of hypotheses are jointly true or not; see, for example, Sankoh et al. (1999, 2003) and Sankoh and Huque (1997). These test methods control the FWER in the weak sense; that is, the control of this error rate is valid only when all null hypotheses included in the intersection are jointly true null hypotheses. Confirmatory clinical trials rarely use these methods for making conclusions for the individual hypotheses; doing so can inflate the FWER, as these methods do not consider all null hypothesis configurations in protecting FWER. These trials, therefore, use methods that control the familywise error rate (FWER) in the strong sense, achieved by controlling this error rate across all null hypothesis configurations. This then assures that the probability of falsely rejecting any true hypothesis, among all the hypotheses tested, is controlled at a pre-specified level α regardless of which and how many other hypotheses are true or false. These strong FWER control methods allow making conclusions for the individual hypotheses without inflating FWER. Methods covered in this chapter are those that fall into this category. However, having said all that, note that intersection hypotheses tests are useful tests and can serve the stated purpose when used in a closed testing scheme, as shown in this chapter.

There is a keen interest in using the Holm and Hochberg methods for testing the primary family of hypotheses of clinical trials, as these methods are more powerful than the Bonferroni procedure. However, these methods allow testing of the secondary family of hypotheses only when all hypotheses in the primary family are first rejected. We have introduced the truncated versions of these methods which have some power advantage and also have some trial α available for the tests of the secondary family of hypotheses, when some (but not all) hypotheses in the primary family are successfully rejected.

Both the gatekeeping and graphical approaches can handle the following two cases: (a) Primary hypotheses test results must not depend in any way on the results of the secondary hypotheses test results; (b) the primary hypotheses, which initially remain unrejected in the testing scheme, can be re-tested on recycling some α from the test results of the secondary hypotheses. This recycling of α in (b) for testing a primary hypothesis obviously makes the significance level for testing this primary hypothesis dependent on the rejections of one or more hypotheses in the secondary

family. In our opinion, such a re-testing of a primary hypothesis, initially not rejected in the first round, is quite valid when at least one primary hypothesis is first rejected so that one is able to characterize a clinically relevant treatment benefit related to that primary hypothesis whose result is independent of the result of the secondary hypotheses tests.

Acknowledgements The authors are grateful to Dr. Lisa LaVange for supporting this chapter. We are also thankful to Drs. Frank Bretz and Dong Xi for providing detailed comments which helped in improving the readability of the materials presented.

Disclaimer This paper reflects the views of the authors and must not be construed to represent FDA's views or policies.

References

Alosh, M., & Huque, M. F. A. (2009). Flexible strategy for testing subgroups and overall populations. *Statistics in Medicine, 28,* 3–23.

Alosh, M., & Huque, M. F. (2010). A consistency-adjusted alpha-adaptive strategy for sequential testing. *Statistics in Medicine, 29,* 1559–1571.

Alosh, M., Bretz, F., & Huque, M. F. (2014). Advanced multiplicity adjustment methods in clinical trials. *Statistics in Medicine, 33*(4), 693–713.

Bauer, P. (1991). Multiple testing in clinical trials. *Statistics in Medicine, 10,* 871–890.

Bauer, P., Rohmel, J., Maurer, W., & Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine, 17,* 2133–2146.

Bretz, F., Maurer, W., Brannath, W., & Posh, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine, 28,* 586–604.

Bretz, F., Maurer, W., & Hommel, G. (2011a). Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in Medicine, 30,* 1489–1501.

Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W., & Rohmeyer, K. (2011b). Graphical approaches for multiple comparison procedures using weighted Bonferroni Simes or parametric tests. *Biometrical Journal, 53*(6), 894–913.

Bretz, F., Maurer, W., & Maca, J. (2014). Graphical approaches to multiple testing. In W. Young & D. G. Chen (Eds.), Chapter 14 in: *Clinical trial biostatistics and biopharmaceutical applications* (pp. 349–394). Boca Raton: Chapman and Hall/CRC press.

Burman, C. F., Sonesson, C., & Guilbaud, O. (2009). A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine, 28,* 739–761.

Dmitrienko, A., Offen, W. W., & Westfall, P. H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine, 22,* 2387–2400.

Dmitrienko, A., Tamhane, A. C., Wang, X., & Chen, X. (2006). Stepwise gatekeeping procedures in clinical trial applications. *Biometrical Journal, 48*(6), 984–991.

Dmitrienko, A., & Tamhane, A. C. (2007). Gatekeeping procedures with clinical trial applications. *Pharmaceutical Statistics, 6,* 171–180.

Dmitrienko, A., Tamhane, A. C., & Wiens, W. (2008). General multi-stage gatekeeping procedures. *Biometrical Journal, 50,* 667–677.

Dmitrienko, A., D'Agostino, R., & Huque, M. F. (2013). Key multiplicity issues in clinical drug development. *Statistics in Medicine, 2013*(32), 1079–1111.

Dmitrienko, A., & Tamhane, A. C. (2009). Gatekeeping procedures in clinical trials. In A. Dmitrienko, A. C. Tamhane, & F. Bretz (Eds.), *Multiple testing problems in pharmaceutical statistics* (Chap. 1). Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series.

Fisher, L. D., & Moyé, L. A. (1999). Carvedilol and the Food and Drug Administration approval process: An introduction. *Controlled Clinical Trials, 20,* 1–15.
Gabriel, K. R. (1969). Simultaneous test procedures—some theory of multiple comparisons. *The Annals of Mathematical Statistics, 40,* 224–520.
Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.
Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75,* 800–802.
Hochberg, E., & Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York: Wiley.
Hommel, G., Bretz, F., & Maurer, W. (2007). Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine, 26,* 4063–4073.
Hsu, J., & Berger, R. L. (1999). Stepwise confidence intervals without multiplicity adjustment for dose response and toxicity studies. *Journal of the American Statistical Association, 94,* 468–482.
Hung, H. M. J., & Wang, S. J. (2009). Some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics, 19,* 1–11.
Hung, H. M. J., & Wang, S. J. (2010). Challenges to multiple testing in clinical trials. *Biometrical Journal, 52,* 747–756.
Huque, M. F., & Alosh, M. (2008). A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *Journal of Statistical Planning and Inference, 138,* 321–335.
Huque, M. F., & Alosh, M. (2012). A consistency-adjusted strategy for accommodating an underpowered primary endpoint. *Journal of Biopharmaceutical Statistics, 22*(1), 160–179.
Huque, M. F., Dmitrienko, A., & D'Agostino, R. (2013). Multiplicity issues in clinical trials with multiple objectives. *Statistics in Biopharmaceutical Research, 5*(4), 321–337.
ICH (International Conference on Harmonization). (1998). *Statistical Principles for Clinical Trials (E-9)*. http://www.fda.gov/cder/guidance/.
Koch, G. G., & Gansky, S. A. (1996). Statistical considerations for multiplicity in confirmatory protocols. *Drug Information Journal, 30,* 523–534.
Li, J., & Mehrotra, D. V. (2008). An efficient method for accommodating potentially under powered primary endpoints. *Statistics in Medicine, 27,* 5377–5391.
Li, J. (2013). Testing each hypothesis marginally at alpha while still controlling FWER: How and when. *Statistics in Medicine, 32,* 1730–1738.
Marcus, R., Peritz, E., & Gabriel, K. R. (1976). On closed testing procedure with special reference to ordered analysis of variance. *Biometrika, 63,* 655–660.
Maurer, W., Hothorn, L. A., & Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In J. Vollman (Ed.), *Biometrie in der chemische-pharmazeutichen Industrie* (Vol. 6). Stuttgart: Fischer Verlag.
Maurer, W., Glimm, E., & Bretz, F. (2011). Multiple and repeated testing of primary, co-primary and secondary hypotheses. *Statistics in Biopharmaceutical Research, 3*(3), 336–352.
Maurer, W., & Bretz, F. (2013). Mutiple testing in group Sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research, 5*(4), 311–320.
O'Neill, R. T. (1997). Secondary endpoints cannot be validly analyzed if primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials, 18,* 550–556.
Rauch, G., Wirths, M., & Keiser, M. (2014). Consistency-adjusted alpha allocation methods for a time-to-event analysis of composite endpoints. *Computational Statistics & Data Analysis, 75,* 151–161.
Sankoh, A. J., D'Agostino, R., & Huque, M. F. (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Statistics in Medicine, 22,* 3133–3150.
Sankoh, A. J., & Huque, M. F. (1997). Some comments on frequently used multiple endpoint adjustment methods I clinical trials. *Statistics in Medicine, 16,* 2529–2542.
Sankoh, A. J., Huque, M. F., Russel, H. K., & D'Agostino, R. (1999). Global two-groupmultiple endpoint adjustment methods in clinical trials. *Drug Information Journal, 33,* 119–140.

Song, Y., & Chi, G. Y. (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine, 26,* 3535–3549.

Westfall, P. H., & Krishen, A. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference, 99,* 25–40.

Wiens, B. L. (2003). A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics, 2,* 211–215.

Wiens, B. L., & Dmitrienko, A. (2005). The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics, 15,* 929–942.

Williams, D. A. (1971). A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, *27*, 103–117. Correction: 31: 1019.

Xi, D., & Tamhane, A. C. (2015). Allocating significance levels in group sequential procedures for multiple endpoints. *Biometrical Journal, 57*(1), 90–107.

Chapter 7
Statistical Testing of Single and Multiple Endpoint Hypotheses in Group Sequential Clinical Trials

Mohammad Huque, Sirisha Mushti and Mohamed Alosh

7.1 Introduction

It is well recognized that a clinical trial of fixed-sample design planned without interim looks can falsely reject a hypothesis of no treatment effect on an endpoint by chance alone. This error commonly known as the false positive error or the Type I error can be excessive if the trial tests more than one hypothesis in the same study. This inflation of the Type I error is of concern as it can lead to false conclusions of treatment benefits in a trial. However, many statistical approaches for confirmatory clinical trials are now available for keeping the probability of falsely rejecting any hypothesis in testing a family of hypotheses (i.e., the familywise Type I error rate) controlled to a specified level; see, for example, a recently released FDA draft guidance "Multiple Endpoints in Clinical Trials," and Alosh et al. (2014).

However, many confirmatory clinical trials accrue patients over many months and enroll hundreds to thousands of patients; this is a widespread practice, for example, for some cardiovascular and oncology trials. Investigators, bound by ethical and economic constraints, usually design these large trials with interim looks, with the possibility of stopping the trial early at an interim stage if the study treatment has the desired efficacy that is clinically relevant, or if it is futile to continue the study, either for lack of efficacy of the study treatment or for safety concerns. These clinical trials are normally recognized as group sequential (GS) clinical trials. The Type I error rate for GS trials, even for the simplest case of testing a single hypothesis, can be inflated

M. Huque (✉)
Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, USA
e-mail: huque.stat@gmail.com

S. Mushti
Division of Biometrics V, Office of Biostatistics, OTS, CDER, FDA, Silver Spring, MD, USA

M. Alosh
Division of Biometrics III, Office of Biostatistics, OTS, CDER, FDA, Silver Spring, MD, USA

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_7

if there are no adjustments for multiple looks, as compared to conventional non-GS trials, because of the repeated tests of the same hypothesis at interim looks. In GS trials, the same hypothesis is tested at different looks as the trial data accumulates over the time course of the trial, until the hypothesis is rejected or the trial reaches the final look for the last test of the hypothesis. Consequently, for assuring the credibility of a treatment benefit result even for a single-hypothesis GS trial, it is considered necessary to use a statistical adjustment method for controlling the probability of a Type I error at a pre-specified level through proper design and analysis methods that are prospectively planned.

There is an extensive literature for GS trials with plans to test a single primary hypothesis of a trial with repeated testing on accumulating data observed at different looks, and to stop the trial early at a look either for efficacy or for futility reasons. This literature covers in detail the technical and operational aspects of such trials, explaining how to plan, conduct, and analyze accumulating data of such trials. Emerson (2007) is an excellent review article on this topic. Also, there are useful books on this topic, including Whitehead (1997), Jennison and Turnbull (2000), and Proschan et al. (2006). Also, there are classical papers on this topic that are of historical importance, such as Armitage et al. (1969), Pocock (1977), O'Brien and Fleming (1979), and Lan and DeMets (1983). In addition, there are some extensions of the methods for multi-arm group sequential trials, e.g., comparison of multiple doses of the same treatment to a common control on a single primary endpoint with interim looks; see, for example, Follmann et al. (1994), Jennison and Turnbull (2000), Hellmich (2001), and Stallard and Friede (2008).

However, modern clinical trials are designed with multiple endpoints; some of these endpoints are given primary and secondary designations. The primary endpoint family along with their hypotheses holds a special position: If the study wins on one or more of its primary endpoint hypotheses then, depending on the level of evidence desired for this win, one can characterize a clinically relevant benefit of the study treatment. In this regard, O'Neill (1997), based on clinical and statistical considerations, made the case that secondary endpoint hypotheses need to be tested only when there is at least one rejection of the primary endpoint hypotheses leading to a clinically relevant benefit of the study treatment. Several innovative statistical procedures for confirmatory clinical trials were proposed that maximize the power for the tests of the primary hypotheses. In doing so, these approaches consider O'Neill's stipulation along with possibility of assigning weights to the different endpoint hypotheses and other logical restrictions. Further, these test procedures control the familywise Type I error rate (FWER) in the "strong sense" (see, e.g., Hochberg and Tamhane 1987), so that the conclusion of treatment efficacy can be made at the individual endpoints or hypotheses levels.

There is a fair amount of literature regarding these novel procedures for fixed-sample clinical trials but not so for GS clinical trials which are frequent for cardiovascular and oncology trials. Examples of such procedures for fixed-sample trial designs include the gatekeeping procedures (see, e.g., Dmitrienko et al. 2003, 2008; Dmitrienko and Tamhane 2009; and Huque et al. 2013 among others) and the graphical procedures (see, e.g., Bretz et al. 2009, 2011, 2014). The development of the

gatekeeping procedures and the graphical method have relied, either explicitly or implicitly, on shortcuts to the closed test procedure, as discussed by Hommel et al. (2007). These developments that utilize short-cut testing have been possible for weighted Bonferroni tests of the intersection hypotheses that satisfy "consonance" property (Hommel et al. 2007). Thereafter, the interest has been as to whether a similar approach for testing multiple hypotheses is possible for GS clinical trials. Recent publications, including Glimm et al. (2010), Tamhane et al. (2010), Maurer and Bretz (2013), Ye et al. (2013), Xi and Tamhane (2015), and Xi et al. (2016), have made this possible and have advanced multiple hypotheses testing methods for GS trials.

Tang and Geller (1999) proposed a general closed testing scheme for testing multiple hypotheses for GS clinical trials. This scheme, though conceptually simple to follow, seems complex to apply in practice, except for certain special situations. By taking advantage of the Hommel et al.'s findings and those of others, we make the case that that Tang and Geller's scheme can be simplified for application purposes by developing short-cut closed test procedures using, for example, the weighted Bonferroni tests. These short-cut procedures for testing multiple hypotheses in GS clinical trials also allow, indirectly, recycling the unused significance level of a rejected hypothesis to testing other hypotheses in a trial.

In this chapter, we first review the classical O'Brien-Fleming (OF) and Pocock (PK) approaches as well as the α-spending function methods, for setting the boundaries in a standard GS clinical trial for repeated testing of a single primary hypothesis. We will call herewith the α-spending function methods as spending function methods. As we will see later, these boundaries computed from the spending function approaches for testing a single hypothesis can still be used for testing multiple hypotheses in GS trials. Consequently, software developed for standard GS trials with a single-hypothesis test can also be used for multiple hypotheses tests. We also touch on the Tang and Geller (1999) closed testing approach as it is of historical importance and show that for testing two primary hypotheses of a trial, this approach simplifies when the weighted Bonferroni test is used for testing the intersection hypothesis. We then visit the graphical approach, for testing multiple primary and secondary hypotheses of GS trials, as discussed by Mauer and Bretz (2013), and present an illustrative example for testing two primary and two secondary endpoints of a trial. Thereafter, we consider the case that when the trial stops after the rejection of a primary hypothesis at a look say for ethical reasons, then other hypotheses need to be tested at the same look, as discussed by Tamhane et al. (2010). We close this chapter with some concluding remarks. Finally, we should point out that in all the discussions and methods presented for deriving boundaries of the GS trials and all tests considered are 1-sided comparing a study treatment to control.

7.2 Testing of a Single Hypothesis in a GS Trial

As in fixed-sample trials, the endpoints in a GS trial can be continuous, binary, or time-to-event. Although the associated test statistics for these endpoints may appear dissimilar, they share a common property: They can be expressed in terms of the standardized sums of independent observations of a random variable. Consequently, they span asymptotically the same joint distribution across time points of multiple looks of the data. Therefore, for the sake of simplicity in this chapter, we assume that the multiple endpoints considered are continuous, and the sample size for each arm of a 2-arm trial designed to compare the study treatment to control remains equal for each endpoint at each look. This case of equal sample size can be easily extended to the case when the sample size for the treated and control arms of the trial at a look can be of different sizes. Also, we consider the case that the total sample size for the final look is fixed in advance. In our discussion of GS trials, we do not consider them adaptive when the investigator continues to modifying the trial design based on the earlier results or what is known as adaptive study design. Adaptive study designs may allow for the possibility of adjusting the sample size of the trial, redefining the endpoint, or modifying the patient population based on the results of an interim look of the data of the trial. Methodological approaches for GS trials with such adaptations are more complex, and some of the assumptions and statements made here may not be valid. With these considerations, we first consider the case of testing a single endpoint hypothesis $H_0 : \delta \leq 0$ against the alternative hypothesis $H_a : \delta > 0$ for a trial with $K - 1$ interim looks and a final look, for a total of $K \geq 2$ looks. A positive value of δ indicates that the test treatment is better than the control.

7.2.1 Test Statistics and Their Distributions

Consider a 2-arm randomized trial designed to compare a treatment with a control on a single primary endpoint based on a total sample size of N subjects per arm. Let S_{n_1} be the sum statistic for the treatment difference at look 1 based on n_1 subjects per treatment arm. This sum statistics at look 1 is the sum of endpoint observations on n_1 subjects in the treatment arm minus the sum of endpoint observations on n_1 subjects in the control arm. Define the B-value at look 1 as

$$B(t_1) = S_{n_1}/\sqrt{V_N}, \text{ where } V_N = \text{Var}(S_N) = 2N\sigma^2. \tag{7.2.1}$$

In (7.2.1), S_N is the sum statistic for the final look yet to be observed and σ^2 is the known variance of individual observations which remains constant throughout the trial regardless of whether the subject observed is in the treatment arm or in the control arm. The value t_1 at look 1, usually known as the information fraction or the information time at look 1, is given by

$$\text{Var}\{B(t_1)\} = n_1/N = t_1. \tag{7.2.2}$$

Note that calling here $n_1/N = t_1$ as the information fraction or information time assumes that the sample sizes for the treatment and control groups are equal at each look and the variance of individual observations remains constant. In general, if d_1 and d_2 denote asymptotically normal estimates of a treatment group difference at interim and final looks, then the information fraction is defined as $I = Var(d_2)/Var(d_1)$. For normal outcomes, information time is the proportion of data available at the interim look, relative to the planned maximum if the trial is not stopped early. However, in presenting our results, for simplicity, we maintain our assumptions of equal sample sizes and constant variance. These results easily extend to the general case (Jennison and Turnbull 2000).

The standardized test statistic $Z(t_1)$ for testing H_0 at look 1 can then be expressed as

$$Z(t_1) = S_{n_1}/\sqrt{V_{n_1}} = (S_{n_1}/\sqrt{V_N})\sqrt{V_N/V_{n_1}} = B(t_1)/\sqrt{t_1}. \tag{7.2.3}$$

The relationship in (7.2.3) follows from $\text{Var}(S_{n_1}) = 2n_1\sigma^2$ and $V_N/V_{n_1} = 1/t_1$. Now consider the second look with the sample size of $n_2 = n_1 + r$ per treatment arm. Then $B(t_2) = (S_{n_1} + S_r)/\sqrt{V_N}$ where S_r is the sum statistic for the treatment difference based on the new data available at look 2. Consequently,

$$\text{Var}\{B(t_2)\} = n_2/N = t_2, \text{Cov}\{B(t_1), B(t_2)\} = t_1,$$

and

$$\text{Corr}\{B(t_1), B(t_2)\} = \text{Corr}\{Z(t_1), Z(t_2)\} = \sqrt{t_1/t_2} \text{ for } t_1 \le t_2. \tag{7.2.4}$$

Given $t_1 \le t_2 \le \cdots \le t_k \le \cdots \le t_K = 1$, we assume that $B(t_1), B(t_2), \ldots, B(t_K)$ follow a multivariate normal distribution with

$$\text{E}\{B(t_k)\} = 0 \text{ under } H_0 \text{ and } \text{Cov}\{B(t_k), B(t_l)\} = t_k \text{ for } t_k \le t_l \le t_K. \tag{7.2.5}$$

Therefore, the normal Z-statistics $\{Z(t_k) = B(t_k)/\sqrt{t_k}\}$ for $k = 1, \ldots, K$ follow a multivariate normal distribution with

$$\text{E}\{Z(t_k)\} = 0 \text{ under } H_0 \text{ and } \text{Cov}\{Z(t_k), Z(t_l)\} = \sqrt{t_k/t_l} \text{ for } t_k \le t_l \le t_K. \tag{7.2.6}$$

The non-central expected value of $B(t_k)$ in terms of the information fraction t_k is given by:

$$\text{E}\{B(t_k)\} = n_k\delta/\sqrt{2N\sigma^2} = (n_k/N)\sqrt{N/2}(\delta/\sigma) = t_k\theta, \tag{7.2.7}$$

where $\theta = \sqrt{N/2}(\delta/\sigma)$ is the "drift parameter." Consequently, the non-central expected value of $\mathrm{E}\{Z(t_k)\} = \sqrt{t_k}\theta$.

Note that $\theta = z_{1-\alpha} + z_{1-\beta}$ for a fixed-sample non-GS trial, where for such a trial, α is the probability of falsely rejecting the null hypothesis $H_0 : \delta \leq 0$ of no treatment effect in favor of the alternative hypothesis $H_a : \delta > 0$ of treatment effect, and power $1-\beta$ is the probability of rejecting H_0 when given the true treatment difference $\delta = \delta_0 > 0$. For example, when the trial $\alpha = 0.025$ and power $1-\beta = 0.90$, then $\theta = 3.2415$. Here the notation z_{1-x} stands for the deviate such that $\Pr(U \leq z_{1-x}) = 1-x$ with $0 \leq x \leq 1$, where U is the normal N(0, 1) random variable. More details about $B(t)$ values and $Z(t)$ normal scores can be found in Proschan et al. (2006) and Lan and Wittes (1988). In the following, we show how the well-known methods by Pocock (1977) and O'Brien and Fleming (1979) rely on these B-values and z-scores in finding their local significance levels, i.e., GS-boundary values, for the repeated testing of H_0. For convenience, we will call these historical methods as PK and OF methods and their boundaries as PK and OF classical boundaries.

7.2.2 Classical PK and OF Boundaries

When analyses of accumulating data of a GS trial occur at equally spaced information times, then the PK boundary is a constant boundary on the z-scale. That is, if $t_k = k/K$ for $k = 1, \ldots, K$, the constant PK boundary $c_{PK}(\alpha, K) = x$ for 1-sided tests can then be obtained by solving for x in the following equation:

$$\Pr[\bigcap_{k=1}^{K} \{Z(t_k) \leq x\} | H_0] = 1 - \alpha \text{ with } t_k = k/K \text{ for } k = 1, \ldots, K, \tag{7.2.8}$$

such that the Type I error rate is controlled at level α. This equation can be solved under the assumption that the joint distribution of the test statistics $\{Z(t_k); k = 1, \ldots, K\}$ is multivariate normal with zero mean vector and correlation matrix $(\rho_{kl}) = \left(\sqrt{t_k/t_l}\right)$ with $t_k \leq t_l$. For example, $c_{PK}(\alpha, K) = 2.28947$ for $K = 3$, ($t_1 = 1/3$, $t_2 = 2/3$, and $t_3 = 1$), and $\alpha = 0.025$. For solving for x in (7.2.8), we wrote SAS/IML codes that calculated the left-hand side of the equation using PROBBNRM and QUAD functions of SAS. PROBBNRM is a SAS function which gives values of the cumulative distribution functions of a standard bivariate normal distribution on specifying the value of the two variables and the correlation coefficient between them. QUAD is a SAS function which integrates numerically a function over an interval. This calculation expressed the joint distribution of $\{Z(t_k); k = 1, 2, 3\}$ as the product of the distribution of $Z(t_1)$ and the conditional bivariate distribution of $Z(t_2)$ and $Z(t_3)$ given $Z(t_1) = z(t_1)$.

Jennison and Turnbull (2000) and Proschan et al. (2006) include 2-sided PK boundary values for different values of K, and $\alpha = 0.01$, 0.05, and 0.10. These 2-sided boundary values at level α, if taken as 1-sided boundary values at level $\alpha/2$,

may not be identical to the actual 1-sided boundary values obtained from (7.2.8); see, for example, Sect. 2.4 in Wassmer and Brannath (2016). The PK boundary values for 2-sided tests are obtained by replacing $Z(t_k) \leq x$ by $|Z(t_k)| \leq x$ in (7.2.8). Thus, a GS trial, designed with PK boundary with looks at equally spaced information times with given α and K, would reject H_0 for efficacy and stop the trial at look k with the information fraction t_k when $Z(t_k) > c_{PK}(\alpha, K)$.

Likewise, the OF boundary is a constant boundary on the B-value scale when the trial looks occur at equally spaced information times. Therefore, when $t_k = k/K$, for $k = 1, \ldots, K$, the 1-sided OF boundary value can be obtained by solving for x in the following equation:

$$\Pr[\bigcap_{k=1}^{K} \{B(t_k) \leq x\} | H_0] = 1 - \alpha \text{ with } t_k = k/K \text{ for } k = 1, \ldots, K.$$

Using $Z(t_k) = B(t_k)/\sqrt{t_k}$ the above equation can be expressed as in (7.2.9) to solve for x using the joint distribution of the test statistics $\{Z(t_k); k = 1, \ldots, K\}$ as a multivariate normal with zero mean vector and correlation matrix $(\rho_{kl}) = \left(\sqrt{t_k/t_l}\right)$ for $t_k \leq t_l$:

$$\Pr[\bigcap_{k=1}^{K} \left\{Z(t_k) \leq x/\sqrt{t_k}\right\} | H_0] = 1 - \alpha \text{ with } t_k = k/K \text{ for } k = 1, \ldots, K. \quad (7.2.9)$$

For example, when $K = 2$, ($t_1 = 1/2$ and $t_2 = 1$), $\alpha = 0.025$, and the tests are 1-sided, then solving the equation PROBBNRM $(x\sqrt{2}, x, \sqrt{1/2}) = 0.975$ gives the value of $x = 1.97742$ which in turn gives the OF boundary values of $c_1(\alpha, K) = x\sqrt{2} = 2.796494$ for the first look at $t_1 = 1/2$ and $c_2(\alpha, K) = x = 1.97742$ for the final look on the z-score scale with the corresponding boundary values of $\alpha_1(\alpha, K) = 0.002583$ and $\alpha_2(\alpha, K) = 0.023997$ on the p-value scale. Thus, if a GS trial is designed with two looks with an interim look at $t_1 = 1/2$, and $\alpha = 0.025$, then H_0 will be rejected when the p-value at this look is less than $\alpha_1(\alpha, K) = 0.002583$ stopping the trial early; otherwise, the trial will continue to the next and final look, and H_0 will be rejected there when the p-value at this look is less than $\alpha_2(\alpha, K) = 0.023997$.

Jennison and Turnbull (2000) and Proschan et al. (2006) provide values of x for 2-sided tests for different values of K and $\alpha = 0.01$, 0.05, and 0.1. These 2-sided boundary values at level α, if read as 1-sided boundary values at level $\alpha/2$, may not agree with the actual 1-sided boundary values. Note that the methods described in this section are of historical importance and are not so frequently used; they lack flexibility because managing analysis at equally spaced information time can be challenging. A more flexible approach for GS trials is the spending function approach described in the next section.

7.2.3 Spending Function Approach

The classical PK and OF boundaries introduced above require specifying the total number of looks at equally spaced information times. This can be inconvenient for clinical trial applications as the Data Safety Monitoring Board (DSMB) or any other group charged with performing interim looks of the accumulating clinical trial data may have to postpone a look for logistical reasons, or may decide to have a look at an unspecified time because of certain concerns. Lan and DeMets (1983) proposed the spending function approach for this and showed that the construction of GS boundaries do not require pre-specification of the number or timings of looks.

Any non-decreasing function $f(\alpha, t)$ in the information time t, over the interval $0 \leq t \leq 1$ and parameterized by the overall significance level α for testing H_0, can be a spending function if it satisfies the following conditions: $f(\alpha, t) \leq f(\alpha, t')$ for $0 \leq t \leq t' \leq 1$; $f(\alpha, t = 0) = 0$; and $f(\alpha, t = 1) = \alpha$. A commonly used spending function for clinical trials is the OF-like:

$$f_1(\alpha, t) = 2\{1 - \Phi(z_{1-\alpha/2}/\sqrt{t})\},$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution.

Note that $f_1(\alpha, 0) = 0$ and $f_1(\alpha, 1) = \alpha$. If the trial had only 2 looks, one at $t = 1/2$ and the other at $t = 1$, and $\alpha = 0.025$, then $f_1(\alpha = 0.025, t = 1/2) = 2(1 - \Phi(2.241403/0.70711)) = 2\{1 - \Phi(3.1698)\} = 0.001525$ and $f_1(\alpha = 0.025, t = 1) = \alpha$. One can then find the significance level x for the final look by solving the equation $\Pr\{(P_1 < 0.001525) \cup (P_2 < x)\} = 0.025$. The next section shows how these equations are solved. The advantage of using the OF-like spending function for clinical trials is its shape which is convex. This allows spending very little of the total α for early looks and saves most of it for latter looks when the trial has sufficient number of patients exposed to the new treatment. The idea is to stop the trial early only when the treatment effect size is sufficiently large and clinically convincing.

Table 7.1 includes a few other spending functions. These and other spending functions give the cumulative Type I error rate spent at look k with the associated information fraction t_k. This cumulative value does not give directly the local significance level $\alpha_k(\alpha, t_k)$ (i.e., the boundary value) for testing H_0 at look k, except when $k = 1$ (the first look). Note that these boundary values are on the p-value scale and need to be converted for presentation on the z-scale. Finding $\alpha_k(\alpha, t_k)$ requires additional calculations which we describe in the following with an example. These calculations usually require solving equations in multiple integrals and are not easy when $K \geq 3$. Special computer software is normally used for this.

Table 7.1 Examples of spending functions

Linear	Pocock-like	Hwang-Shi-Decani (1990)
$f_2(\alpha, t) = \alpha t$	$f_3(\alpha, t) = \alpha \log_e\{1 + (e-1)t\}$	$f_4(\alpha, t) = \alpha\big[\{1 - \exp(-\lambda t)\}/\{1 - \exp(-\lambda)\}\big]$, for $\lambda \neq 0$

7.2.4 Calculations of Boundary Values Using Spending Functions

We illustrate the use of spending functions for finding the local significance level $\alpha_k(\alpha, t_k)$ at look k with the information fraction t_k, so that H_0 will be rejected when the 1-sided p-value p_k at this look is less than $\alpha_k(\alpha, t_k)$. Suppose a trial uses the OF-like spending function to control the Type I error rate at level $\alpha = 0.025$. Suppose that the first look occurs at $t_1 = 0.30$. Then at this look, we spend

$$f_1(\alpha = 0.025, t_1 = 0.30) = 2\left\{1 - \Phi\left(z_{1-\alpha/2}/\sqrt{0.30}\right)\right\}$$
$$= 2\left\{1 - \Phi\left(\frac{2.2414027}{\sqrt{0.30}}\right)\right\} = 0.0000427$$

Therefore, at this look, $\alpha_1(\alpha, t_1) = 0.0000427$ and the critical value $c_1(\alpha, t_1) = 3.9285725$ from $\Pr\{Z(t_1) > c_1(\alpha, t_1)\} = 0.0000427$; one will reject H_0 and stop the trial at the first look if $p_1 < 0.0000427$ or $Z(t_1) > 3.9285725$. Thus, at this look the investigator spends very little of the total $\alpha = 0.025$.

Suppose that the trial did not stop at the first look and the investigator decides to have the second look at $t_2 = 0.65$. Then the cumulative alpha spent at this look is

$$f_1(\alpha = 0.025, t = 0.65) = 2\left\{1 - \Phi\left(z_{1-\alpha/2}/\sqrt{0.65}\right)\right\}$$
$$= 2\left\{1 - \Phi\left(\frac{2.2414027}{\sqrt{0.65}}\right)\right\} = 0.0054339$$

Therefore, we determine the boundary critical values of $c_2(\alpha, t_2) = 2.5479$ or $\alpha_2(\alpha, t_2) = 0.0054187$ by solving the equation: $\Pr[\{(Z(t_1) > 3.9285725\} \cup \{(Z(t_2) > c_2(\alpha, t_2)\}] = 0.0054339$. Therefore, one can reject H_0 at the second look and stop the trial, if at this look, the observed p-value $p_2 < 0.005187$ or $Z(t_2) > 2.5479$.

Suppose the trial did not stop at this second look and the investigator moves to the final look at $t_3 = 1$. Then the cumulative alpha spent at the final look is $\alpha = 0.025$. One can then find $c_3(\alpha, t_3)$ by solving the equation:

$$\Pr[\{(Z(t_1) > 3.9285725\} \cup \{(Z(t_2) > 2.5479\} \cup \{(Z(t_3) > c_3(\alpha, t_3)\}] = 0.025$$

Table 7.2 Examples for the OF-like spending function with $\alpha = 0.025, 0.0125$, $K = 3$, and 1-sided tests

Look #	Information fraction	Cumulative α spent	Boundaries
$\alpha = 0.025$			
1	0.30	0.00004	0.00004
2	0.65	0.00543	0.00542
3	1.00	0.025	0.02331
$\alpha = 0.0125$			
1	0.30	0.00001	0.00001
2	0.65	0.00194	0.00194
3	1.00	0.0125	0.01188

Solving this equation gives $c_3(\alpha, t_3) = 1.9897$ and $\alpha_3(\alpha, t_3) = 0.023312$. Therefore, one can reject H_0 at the final look if at this look the p-value $p_3 < 0.023312$ or $Z(t_3) > 1.9897$.

A general recursive equation for finding $c_k(\alpha, t_k)$ and $\alpha_k(\alpha, t_k)$ for a spending function $f(\alpha, t)$ is given by $f(\alpha, t_k) = f(\alpha, t_{k-1}) + \Pr\left[\{\bigcap_{i=1}^{k-1} Z(t_i) \le c_i(\alpha, t_i)\} \cap \{Z(t_k) > c_k(\alpha, t_k)\}\right]$ for $k \ge 2$. There are software available that give values of $c_k(\alpha, t_k)$ and $\alpha_k(\alpha, t_k)$ for OF-like and other spending functions, see Zhu et al. (2011) for a review of these software. Table 7.2 shows the results from such a software. We show in Sect. 7.3 that such boundaries can also be used for testing multiple hypotheses of GS trials.

7.3 Testing of Multiple Hypotheses in GS Trials

Many GS trials are designed for testing multiple endpoint hypotheses, frequently, for testing two endpoint hypotheses. Two situations generally arise. Consider, for example, a GS trial for testing two endpoint hypotheses. The first case arises when after the rejection of one of the two hypotheses at an interim look the trial does not stop but continues to later looks for testing the other hypothesis. The second case arises when the two hypotheses are hierarchically ordered, e.g., one is primary and the other is secondary. The first hypothesis in the hierarchy (i.e., the primary hypothesis) is allocated first using the full trial α (e.g., $\alpha = 0.025$). If this hypothesis is rejected at an interim look, then the trial stops because of ethical considerations. For example, if the first hypothesis is associated with the mortality endpoint and the second hypothesis with a quality of life measure, then if the trial wins at a look for the mortality endpoint then the trial would generally discontinue for ethical reasons. In that case, the second hypothesis (i.e., the secondary hypothesis) is tested at the same look at which the first hypothesis was rejected. The remainder of this section considers the first case and Sect. 7.4 considers the second case. In the following, we first address methods based on the Bonferroni inequality and then move on to

α-recycling approaches based on the closed testing principle (CTP) of Marcus et al. (1976), and finally to the more recent graphical approach of Maurer and Bretz (2013).

7.3.1 Methods Based on the Bonferroni Inequality

Consider, for example, a trial which for the demonstration of superiority of a new treatment to control specifies two null hypotheses: H_1 and H_2. Rejection of either of the two hypotheses at a look can establish efficacy of the new treatment. However, if the trial rejects one of the two hypotheses at an interim look, the trial can continue to later looks for testing the other hypothesis. For such a trial, the use of the Bonferroni inequality leads to two approaches for a stronger claim. The first approach splits the significance level α as $\alpha_1 + \alpha_2 \leq \alpha$ for testing H_1 at level α_1 and H_2 at level α_2. For example, it may assign $\alpha_1 = 0.005$ for testing H_1 and $\alpha_2 = 0.02$ for testing H_2 for controlling the overall Type I error rate at $\alpha = 0.025$. Tests for H_1 and H_2 can then separately follow in a univariate GS testing framework for the separate control of the Type I error rates at levels α_1 and α_2, respectively, using the same or different spending functions for each. In Sect. 7.3.2, we show that this approach extends to an α-recycling approach, such that, if one of the multiple hypotheses is rejected at a look then the boundary value for testing other hypotheses is updated to larger values.

The second approach uses the Bonferroni inequality differently. It specifies the rejection boundary values as $\alpha'_k(t_k) > 0$ for looks $k = 1, \ldots, K$ such that $\sum_{k=1}^{K} \alpha'_k(t_k) = \alpha$. It then applies a conventional multiple hypothesis testing method at a look for the control of the Type I error rate at the local level $\alpha'_k(t_k)$ at that look. Suppose that $K = 2$, i.e., the trial is designed with two looks, and $\alpha'_1(t_1) = 0.005$ and $\alpha'_2(t_2 = 1) = 0.02$, for the first and second looks, respectively. One can then apply, for example, the conventional Hochberg procedure (1988) for testing H_1 and H_2 at level 0.005 at the first look, and similarly, can apply the same procedure for testing these hypotheses at the final look at level 0.02. The methods discussed in this section for testing two hypotheses generalize to testing more than two hypotheses.

7.3.2 Method Based on the Closed Testing Principle

The closed testing principle of Marcus et al. (1976) provides a general framework for constructing powerful closed test procedures (CTPs) for testing individual hypotheses based on tests of intersection hypotheses of different orders. One starts with a family of individual hypotheses $H_1, \ldots, H_h$ and constructs a closed set $\tilde{\boldsymbol{H}}$ of $2^h - 1$ non-empty intersection hypotheses as follows:

$$\tilde{\boldsymbol{H}} = \left\{ H_{\boldsymbol{J}} = \bigcap\nolimits_{j \in \boldsymbol{J}} H_j, \quad \boldsymbol{J} \subseteq \boldsymbol{I} = \{1, \ldots, h\} \right\}.$$

One then performs an α-level test for each hypothesis $H_{\boldsymbol{J}}$ in $\tilde{\boldsymbol{H}}$ by using, for example, the weighted Bonferroni test. One then rejects an individual hypothesis H_j when all $H_{\boldsymbol{J}}$ for $j \in \boldsymbol{J}$ are rejected by their corresponding α-level tests.

For example, when $h = 2$, the closed set $\tilde{\boldsymbol{H}} = \{H_{12}, H_1, H_2\}$. A CTP will reject the individual hypothesis H_1 only when H_1 and H_{12} are both rejected, each by an α-level test. If one uses, for example, the weighted Bonferroni test for H_{12}, then the procedure cuts down the extra step of testing H_1 after rejecting H_{12}. The weighted Bonferroni test rejects H_{12}, when $p_j < w_j\alpha$ for at least one $j \in \{1, 2\}$, where w_1 and w_2 are the nonnegative weights assigned to H_1 and H_2, respectively, such that $w_1 + w_2 \leq 1$, and p_j are the observed p-values associated with H_j for $j \in \{1, 2\}$. Suppose that this test rejects H_{12} for $j = 1$ on observing $p_1 < w_1\alpha$, then H_1 is automatically rejected, as the significance level α for the test of H_1 satisfies $\alpha \geq w_1\alpha$. This property in its general form, known as the *consonance* property, when satisfied for testing intersection hypotheses in a closed testing procedure, leads to short-cuts of closed test procedures and allows recycling of the significance level of a rejected hypothesis to other hypotheses (Hommel et al. 2007). This property basically means that the rejection of an intersection hypothesis $H_{\boldsymbol{J}}$ by an α-level test implies the rejection of at least one individual hypothesis H_j for $j \in \boldsymbol{J}$.

As a numerical example, consider testing the two hypotheses H_1 and H_2 with $\alpha = 0.025$, and suppose that weights assigned to H_1 and H_2 are $w_1 = 0.8$ and $w_2 = 0.2$, respectively, so that $w_1 + w_2 = 1$. Further, supposed that the associated observed p-values for the tests of H_1 and H_2 were $p_1 = 0.024$ for H_1 and $p_2 = 0.004$ for H_2. The simple weighted Bonferroni test would reject only H_2, as $p_1 > w_1\alpha = 0.020$ and $p_2 < w_2\alpha = 0.005$. However, the weighted Bonferroni based CTP with these weights would reject both hypotheses. This CTP, in its initial step, would reject the intersection hypothesis H_{12} as $p_j < w_j\alpha$ for $j = 2$. Consequently, as the procedure assigns the weights of one for testing each singleton hypotheses, satisfying consonance, it would then reject each of the two hypotheses as $p_j < 1.\alpha = 0.025$ for each $j \in \{1, 2\}$.

In the following, we first visit the GS closed test procedure by Tang and Geller (1999) for testing multiple hypotheses and show that this procedure leads to α-recycling procedures by using weighted Bonferroni tests of intersection hypotheses that satisfy consonance. The Tang and Geller procedure is of historical importance with respect to using the closed testing procedure for testing multiple hypotheses in group sequential trials. Although the procedure sounds complicated in its original form, it can be simplified if the weighted Bonferroni tests, with weights satisfying the consonance property, are used for testing its intersection hypotheses. However, selection of such weights can be cumbersome for testing more than three hypotheses. Section 7.3.3 toward the end illustrates how to find these weights when testing two primary hypotheses and a secondary hypothesis. In general, the graphical approach (Sect. 7.3.5) in this regard is easier to use when testing multiple hypotheses.

Consider testing $h \geq 2$ endpoint hypotheses in a GS trial designed to compare a new treatment to control. Consider, as before, the intersection hypotheses $H_{\boldsymbol{J}}$ for $\boldsymbol{J} \subseteq \boldsymbol{I} = \{1, \ldots, h\}$, i.e., the new treatment to control treatment difference $\delta_j \leq 0$ for all endpoints $j \in \boldsymbol{J} \subseteq \boldsymbol{I}$. Also, consider that multiple looks for the trial occur at

different information times $t \in \{t_1, t_2, \ldots, t_K\}$ such that $t_1 \leq t_2 \leq \cdots \leq t_K = 1$. Let $Z_{\boldsymbol{J}}$ be a test statistic for testing $H_{\boldsymbol{J}}$ (e.g., by a weighted Bonferroni test) and let $Z_{\boldsymbol{J}}(t)$ be the test statistic value of $Z_{\boldsymbol{J}}$ at a look with information fraction t. Further, let $c_{\boldsymbol{J}}(t)$ be the critical value for performing an α-level test of $H_{\boldsymbol{J}}$ at this look by using $Z_{\boldsymbol{J}}(t)$. That is, for each $\boldsymbol{J} \subseteq \boldsymbol{I}$, the $c_{\boldsymbol{J}}(t)$ values for different t (at which times repeated tests occur) satisfy $\Pr\{Z_{\boldsymbol{J}}(t) > c_{\boldsymbol{J}}(t) \text{ for some } t | H_{\boldsymbol{J}}\} \leq \alpha$. Then a closed test procedure for GS trials as proposed by Tang and Geller (1999) can be stated as follows:

Step 1: Start testing $H_{\boldsymbol{I}}$ as in a univariate case of a GS trial but using the group sequential boundary values $c_{\boldsymbol{I}}(t)$ for the test statistics $Z_{\boldsymbol{I}}(t)$, where $\boldsymbol{I} = \{1, \ldots, h\}$.

Step 2: Suppose that $H_{\boldsymbol{I}}$ is rejected first time at the look with $t = t^*$. Then, for rejecting at least one individual hypothesis at this look, apply a CTP to test $H_{\boldsymbol{J}}$ with $\boldsymbol{J} \subseteq \boldsymbol{I}$ using $Z_{\boldsymbol{J}}(t^*)$ and its critical value $c_{\boldsymbol{J}}(t^*)$. Note that $c_{\boldsymbol{J}}(t^*)$ can be different for different $H_{\boldsymbol{J}}$'s. In applying this CTP at $t = t^*$ either (a) none of the individual hypotheses will be rejected, or (b) at least one individual hypothesis H_j will be rejected for $j \in \boldsymbol{I}$.

Step 3(a): In *Step* 2, if none of the individual hypotheses are rejected at $t = t^*$ then continue to the next look; however, if $t^* = 1$ and none of the individual hypotheses are rejected, the trial will stop without the rejection of any hypothesis.

Step 3(b): In *Step* 2, if at least one hypothesis is rejected at $t = t^*$, then exclude the indices of the rejected hypotheses from the index set $\boldsymbol{I}$. With this updated index set $\boldsymbol{I}$, continue to the next look and repeat *Step* 1 and *Step* 2. Note that in this process, all previously rejected hypotheses are assumed rejected at later looks and are removed for further testing.

Step 4: Reiterate the above steps until all hypotheses are rejected or the trial reaches the final look.

Implementing the Tang and Geller (1999) approach for the general case can be complicated because of the computational difficulties in finding $c_{\boldsymbol{J}}(t)$ values for testing $H_{\boldsymbol{J}}$ for different $\boldsymbol{J}$ and different looks. However, this approach simplifies on using univariate tests for $H_{\boldsymbol{J}}$ that satisfy consonance. Examples, of such tests, are the max-T or min-p test, and the un-weighted Bonferroni test. Weighted Bonferroni test which is more useful for clinical trial applications also serves this purpose, but the weights for the weighted Bonferroni tests need to be pre-selected to satisfy consonance. This may be difficult when testing more than three hypotheses. An alternative to this which does not have this issue is the graphical approach addressed in Sect. 7.3.4. The following, however, addresses the weighted Bonferroni test approach and illustrates its application for testing two hypotheses in a GS trial.

In the weighted Bonferroni test approach, to satisfy consonance for the tests of $H_{\boldsymbol{J}}$ for $\boldsymbol{J} \subseteq \boldsymbol{I}$, one pre-selects weights $w_j(\boldsymbol{J})$ for $j \in \boldsymbol{J}$ with $\sum_{j \in \boldsymbol{J}} w_j(\boldsymbol{J}) \leq 1$ so that $w_j(\boldsymbol{J}^*) \geq w_j(\boldsymbol{J})$ for every $\boldsymbol{J}^* \subseteq \boldsymbol{J}$. For these cases, standard software developed for testing a single hypothesis with a spending function approach can still be used for testing multiple hypotheses. The following is an illustrative example for testing two hypotheses H_1 and H_2 in a GS trial.

In the case of testing two hypotheses, a CTP considers a single intersection hypothesis $H_{\boldsymbol{J}}$ with $\boldsymbol{J} = \{1, 2\}$, written as H_{12}, and two individual hypotheses H_1 and H_2. Suppose that for testing H_{12} one assigns weights $w_1\{1, 2\} = 0.8$ and $w_2\{1, 2\} = 0.2$ so that $w_1\{1, 2\}\alpha = 0.02$ and $w_2\{1, 2\}\alpha = 0.005$ with the trial $\alpha = 0.025$. Consonance is satisfied, because after H_{12} is rejected, the weights for testing each of the two individual hypotheses in the CTP is one. The following illustrates how one will test H_1 and H_2 in a GS trial with such initial weights.

Tests at the First Look

Suppose that the first look for the trial occurs at $t = t_1 = 0.30$, and suppose that at this look the unadjusted p-values associated with H_1 and H_2 are $p_1(t_1)$ and $p_2(t_1)$, respectively. The CTP will reject H_{12} by the weighted Bonferroni test if either $p_1(t_1) < \alpha_1(w_1\{1, 2\}\alpha = 0.02, t_1 = 0.30) = \alpha_1(0.02, t_1 = 0.30)$ or $p_2(t_1) < \alpha_2(0.005, t_1 = 0.30)$, where these boundary critical values can be obtained by specifying spending functions f_1 and f_2. If f_1 and f_2 are each OF-like, then

$$\alpha_1(0.020, t_1 = 0.30) = f_1(w_1\{1, 2\}\alpha = 0.02, t_1 = 0.30) = 0.00002$$
$$\alpha_2(0.005, t_1 = 0.30) = f_2(w_2\{1, 2\}\alpha = 0.005, t_1 = 0.30) = 2.977\text{E} - 07$$

Suppose that H_{12} is not rejected at this look with $t_1 = 0.30$ and the trial continues to the second look.

Tests at the Second Look

Suppose that the second look occurs at $t_2 = 0.65$. Further, suppose that at this look the unadjusted p-values associated with H_1 and H_2 are $p_1(t_2)$ and $p_2(t_2)$, respectively. Consequently, the CTP will reject H_{12} at this look if either $p_1(t_2) < \alpha_1(0.02, t_2 = 0.65)$ or $p_2(t_2) < \alpha_1(0.005, t_2 = 0.65)$. The use of the spending functions f_1 and f_2 as OF-like for this look gives the boundary values

$$\alpha_1(0.020, t_1 = 0.65) = 0.0039 \text{ and } \alpha_2(0.005, t_1 = 0.65) = 0.000498.$$

Section 7.2.4 has addressed how these boundary values are calculated. As indicated before, computer software is used to calculate such boundary values.

Now, suppose that $p_2(t_2) < 0.000498$, then H_{12} will be rejected leading to the automatic rejection of H_2 because of the consonance condition being satisfied. Therefore, as H_{12} and H_2 are rejected at $t^* = t_2 = 0.65$, the CTP will test the remaining hypothesis H_1 at the same look with ($t^* = t_2 = 0.65$) with the updated boundary value of $\alpha_1(0.025, t_2 = 0.65) = 0.00542$ by the same OF-like spending function. Thus, there is a recycling of alpha of 0.005 form the rejected H_2 to H_1, updating the alpha of 0.02 to $0.02 + 0.005 = 0.025$ which is incorporated in the first argument of $\alpha_1(0.025, t_2 = 0.65)$. Thus, a CTP with consonance allows recycling of alpha for GS trials, but here, this recycling updates the boundary values for testing H_1 starting from at $t^* = t_2 = 0.65$ using a spending function. Suppose that $p_1(t_2) = 0.015$ which is greater than 0.00542, then H_1 at this second look remains not rejected. The trial then continues to the final look with $t_3 = 1$ for testing H_1.

Test at the Final Look
The final look occurs with $t_3 = 1$ for testing H_1 with the assumption that H_2 (which was rejected at the second look) remain rejected at this look. Therefore, H_1 would be tested at this look at level $\alpha_1(0.025, t_3 = 1) = 0.02331$ by the same OF-like spending function.

7.3.3 Some Key Considerations and Comments

For applications, the spending functions to be used for testing different hypotheses need to be pre-specified, and for interpreting study findings, it is good practice to use the same spending functions for testing different hypotheses. It should be noted that although the total number of looks may not be pre-specified, however, specifying it may help reducing concerns about unnecessary looks of the data. In addition, in our previous discussion, including the illustrative example in Sect. 7.3.2, we assumed that information fractions for the two endpoints are equal at each look. This can be the case for continuous or binary endpoints; however, this may be not the general case. That is, if $t_k(E_1)$ and $t_k(E_2)$ are information fraction for two endpoints at looks $k = 1, .., K$ then it is possible that $t_k(E_1) \neq t_k(E_2)$ for at least one k. This can occur, for example, when E_1 or E_2 are time-to-event endpoints; it may also occur for other situations. Then the question may arise as how to adopt the above procedure for this general case.

In this regard, we note that the above procedure can be easily adopted to address this general case. To illustrate, suppose that in the above example, at the first look $t_1(E_1) = t_1(E_2) = 0.30$, but at the second look $t_2(E_1) = 0.40$ and $t_2(E_2) = 0.65$ and assume that H_{12} is not rejected at the first look; yet, it can be rejected at the second look if either $p_1(t_2) < \alpha_1(0.02, t_2(E_1) = 0.40)$ or $p_2(t_2) < \alpha_1(0.005, t_2(E_2) = 0.65)$. Now, suppose that at this stage H_{12} is rejected by observing that $p_2(t_2) < \alpha_1(0.005, t_2(E_2) = 0.65)$, leading to the rejection of H_2 as before. Therefore, the alpha of 0.005 for the rejected H_2 will now be recycled for testing H_1, that is by updating the old boundary value of $\alpha_1(0.02, t_2(E_1) = 0.40)$ to a new boundary value $\alpha_1(0.025, t_2(E_1) = 0.40)$ at this second look, and to $\alpha_1(0.025, t_3(E_1) = 1)$ at the final look.

Note that in above after rejecting H_2 at the second look, the significance level for testing for H_1 is $\alpha_1(0.025, t_2(E_1) = 0.40)$ which is not equal to $\alpha = 0.025$. Wrongfully, testing H_1 at $\alpha = 0.025$ instead of testing it at level $\alpha_1(0.025, t_2(E_1) = 0.40)$ after the rejection of H_2 can inflate the overall Type I error rate. Also, if the trial stops at a look after rejecting a hypothesis for ethical reasons, say after the rejection of H_2, then one cannot test a second hypothesis such as H_1 at the full significance level of $\alpha = 0.025$. Doing this can inflate the overall Type I error rate, except for the special case when the test statistics for the two hypotheses are independent. We consider this type of GS trials in Sect. 7.4.

The spending functions used to test each hypothesis needs to satisfy a monotonicity property. That is, the difference function $f(\lambda, t_k) - f(\lambda, t_{k-1})$ is monotonically

non-decreasing in λ for $k = 1, \ldots, K$. For example, the OF-like α-spending function satisfies this condition for $\lambda < 0.318$ (Maurer and Bretz 2013).

The above weighted Bonferroni-based CTP for testing two hypotheses can be extended to testing more than two hypotheses if weights assigned for testing intersection hypotheses in a CTP are such that consonance property is guaranteed, that is, weights assigned are such that rejection of an intersection hypothesis in the CTP leads to the rejection of at least one individual hypothesis in that intersection hypothesis. For example, for testing two primary hypotheses H_1 and H_2 and a secondary hypothesis H_3 of a trial, the CTP would consider four intersection hypotheses H_{123}, H_{12}, H_{13} and H_{23} and three individual hypotheses.

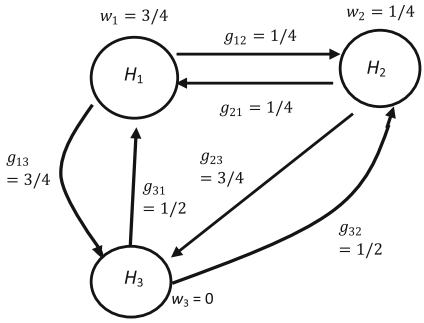
The following selection of weights for performing Bonferroni-based tests of intersection hypotheses in the CTP would then satisfy consonance property. Assign non-negative weights of w_1, w_2, and w_3 associated with indices (1, 2, and 3) of H_{123} to test this hypothesis with $w_1 + w_2 = 1$ and $w_3 = 0$; the selection of $w_3 = 0$ indicates that H_3 is tested only after at least one of the two primary hypotheses is first rejected. Assign weights of $\{w_1, w_2\}$ to H_1 and H_2, respectively to test H_{12}. Similarly, weights of $\{w_1 + \delta_2 w_2, (1 - \delta_2)w_2\}$ to test H_{13}, and weights of $\{w_2 + \delta_1 w_1, (1 - \delta_1)w_1\}$ to test H_{23}, where $0 \leq \delta_1 \leq 1$ and $0 \leq \delta_2 \leq 1$. The weights assigned to each of the individual hypotheses will be one. The selection of these weight and the recycling parameter δ_1 and δ_2, for example, can be based on the trial objectives. Once such weights for performing the weighted Bonferroni tests satisfy consonance, a CTP for testing the above three hypotheses in a GS trial can be proposed.

GS trials that are not properly conducted have the potential of unblinding the trial prematurely, and consequently, this may impact the integrity of the trial and its results. To address this important issue, usually an Independent Data Monitoring Committees (DMC) along with a charter is setup for GS trials. As our focus for this chapter is to overview the general multiple testing approaches for group sequential trials, we do not discuss this issue here. The interested reader may consult relevant literature in this regard, see, e.g., Ellenberg et al. (2017). The concerns about potential unblinding for testing single hypothesis over the course of GS trials remain the same for GS trials with testing multiple hypotheses related to multiple endpoints.

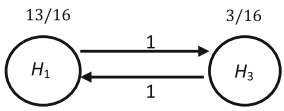
For a GS trial that include testing of multiple hypotheses, a Statistical Analysis Plan (SAP) that explains in sufficient details the design, the analyses method, and the DMC charter, is essential for proper interpretation of study findings. Such a SAP should in general be developed a priori and agreed upon by those involved before launching the trial.

7.3.4 Graphical Approach

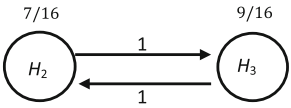
The above weighted Bonferroni-based CTP for testing multiple hypotheses of a GS trial, though possible, can be challenging in finding appropriate weights that guarantee consonance when the number of hypotheses tested are more than a few. The graphical approach of Bretz et al. (2009) which includes a special algorithm for



(a) Initial Graph for testing H_1, H_2, and H_3 along with their weights



(b) Graph after rejecting H_2 in (a), for testing H_1 and H_3 along with their weights



(c) Graph after rejecting H_1 in (a), for testing H_2 and H_3 along with their weights

Fig. 7.1 Graphical representation of testing with two primary hypotheses H_1 and H_2, and one secondary hypothesis H_3

doing this solves this problem. In this approach, one can graphically visualize the weighted Bonferroni tests for multiple hypotheses along with an α-propagation rule by which the procedure recycles the significance level of a rejected hypothesis to other remaining unrejected hypotheses. This graphical approach, originally developed for testing multiple hypotheses of non-GS trials, can also be conveniently used for testing multiple hypotheses of GS trials; see, for example, Maurer and Bretz (2013). The following explains the key concepts of this approach for testing multiple hypotheses.

In this graphical approach, the h individual hypotheses are represented initially by a set of h nodes with nonnegative weight of w_i at node $i (i = 1, \ldots, h)$ such that $\sum_{i=1}^{h} w_i \leq 1$. These weights when multiplied by α represent the local significance levels at those respective nodes. The weight g_{ij} (with $0 \leq g_{ij} \leq 1$) associated with a directed edge connecting the node i to the node j indicates the fraction of the local significance level at the tail node i that is added to the significance level at the terminal node j, if the hypothesis at the tail node i is rejected. For convenience, we will call these directed edges as "arrows" running from one node to the other, and the weight g_{ij} as the "transition weight" on the arrow running from node i to node j.

Figure 7.1 illustrates key concepts of this graphical approach for testing two primary hypotheses H_1 and H_2 and a secondary hypothesis H_3 of a trial. In this figure, the initial Graph (a) shows three nodes. Two nodes represent H_1 and H_2 with weights $w_1 = 3/4$ and $w_2 = (1 - w_1) = 1/4$, respectively. The node for H_3 shows a weight $w_3 = 0$, which can increase only after the rejection of a primary hypothesis. The nonnegative number $g_{12} = 1/4$ is the transition weight on the arrow going from H_1 to H_2; similarly, $g_{21} = 1/4$ is the transition weight on the arrow going from H_2 to H_1. The transition weight on the arrow going from H_1 to H_3 is 3/4 and that on the arrow going from H_2 to H_3 is also 3/4 satisfying the condition that sum of the transition weights of all outgoing arrows from a single node must be bounded above by 1.

Graph (b) of Fig. 7.1 represents the resulting graph after H_2 is rejected in Graph (a). The rejection of this hypothesis frees its weight w_2 which is then recycled to H_1 and H_3 according to an α-propagation rule addressed in the following for the general case. This rule also calculates new transition weights going from one node to the other for the new graph. Graph (c) of Fig. 7.1 similarly shows the resulting graph if H_1 is rejected in Graph (a). The following shows the general graphical procedure for testing h individual hypotheses $H_1, \ldots, H_h$ for a non-GS trial given their individual unadjusted p-values p_j for $j = 1, \ldots, h$.

(0) Set $\boldsymbol{I} = \{1, \ldots, h\}$. The set of weights $\{w_j(\boldsymbol{I}), j \in \boldsymbol{I}\}$ are such that $0 \leq w_j(\boldsymbol{I}) \leq 1$ with the sum $\sum_{j \in \boldsymbol{I}} w_j(\boldsymbol{I}) \leq 1$.

(i) Select a $j \in \boldsymbol{I}$ such that $p_j < \{w_j(\boldsymbol{I})\}\alpha$ and reject H_j; otherwise stop.

(ii) Update the graph as:

(a) $\boldsymbol{I} = \boldsymbol{I} \backslash \{j\}$, i.e., the index set $\boldsymbol{I}$ without the index j

(b)

$$w_l(\boldsymbol{I}) = w_l(\boldsymbol{I}) + w_j(\boldsymbol{I}) g_{jl}, \, l \in \boldsymbol{I}; \; 0, \text{ otherwise} \tag{7.3.1}$$

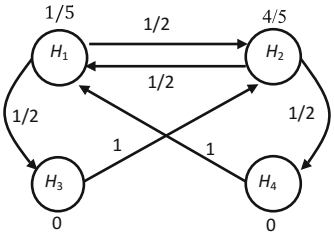
(c)

$$g_{lk} = \frac{g_{lk} + g_{lj} g_{jk}}{1 - g_{lj} g_{jl}}, \text{ where } (l, k) \in \boldsymbol{I}, \, l \neq k \text{ and } g_{lj} g_{jl} < 1; \; 0, \text{ otherwise} \tag{7.3.2}$$

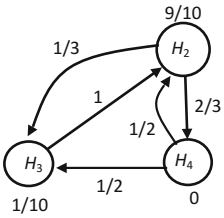
(iii) If $|\boldsymbol{I}| \geq 1$ then go to step (i); otherwise stop

After rejecting H_j, the Eq. (7.3.1) for a new graph updates the weight for H_l to a new weight which is its old weight $w_l(\boldsymbol{I})$ plus the weight $w_j(\boldsymbol{I})$ at H_j multiplied by the transition weight g_{jl} on the arrow connecting H_j to H_l. Also, the transition weights g_{lk} for the new graph are obtained by the algorithm (7.3.2) whose numerator $g_{lk} + g_{lj} g_{jk}$ is the transition weight on the arrow connecting H_l to H_k plus the product of the transition weights on arrows going from H_l to H_k through the rejected hypothesis H_j. The term $g_{lj} g_{jl}$ in (7.3.2) is the product of transition weights on arrows connecting H_l to H_j and then returning to H_l. The approach produces weights $w_l(\boldsymbol{I})$ which satisfy consonance.

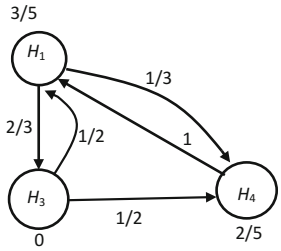
For explaining this procedure, consider a trial, which for demonstrating superiority of a new treatment A + Standard of Care (SOC) to placebo +SOC, plans to test two primary hypotheses H_1 and H_2 and two secondary hypotheses H_3 and H_4, where the pairs (H_1, H_3) and (H_2, H_4) being considered as parent–descendant (Maurer et al. 2011). That is, H_3 is tested only when H_1 is rejected, and similarly, H_4 is tested only when H_2 is rejected. Suppose that the trial specifies a graphical test strategy as in Fig. 7.2 for testing these four hypotheses. The initial Graph (a) in Fig. 7.2 gives a smaller weight of $w_1 = 1/5$ to H_1 as compared to a weight of $w_2 = 4/5$ to H_2 based on the prior experience that the trial may win easily for H_1 at the significance level of $w_1\alpha = 0.005$, but the trial may require a larger significance level of $w_2\alpha = 0.02$ for winning for H_2. As stated before, we assume that all tests in the procedure are



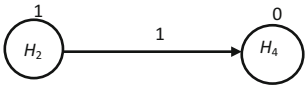
(a) Initial Graph for testing H_1, H_2, H_3, and H_4 along with their weights



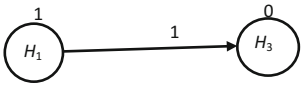
(b) Graph for testing for testing H_2, H_3, and H_4 along with their weights after the rejection of H_1 in (a)



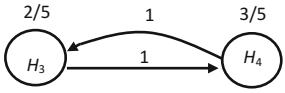
(c) Graph for testing H_1, H_3, and H_4 along with their weights after the rejection of H_2 in (a)



(d) Graph for testing H_2 and H_4 along with their B-weights after the rejection of H_3 in (b)



(e) Graph for testing H_1 and H_3 along with their weights after the rejection of H_4 in (c)



(f) Graph for testing H_3 and H_4 along with their weights after the rejection of H_2 in (b) or H_1 in (c)

Fig. 7.2 Graphical test procedure for two primary hypotheses H_1 and H_2, and two secondary hypotheses H_3 and H_4, where pairs (H_1, H_3) and (H_2, H_4) are parent–descendant

1-sided and the control of the overall Type I error rate is at level $\alpha = 0.025$. The Graph (a) assigns zero-weights to the two secondary hypotheses indicating that we do not want to reject a secondary hypothesis until its parent primary hypothesis is first rejected.

In Graph (a) of Fig. 7.2, $g_{12} = g_{21} = g_{13} = g_{24} = 1/2$ and $g_{32} = g_{41} = 1$. These settings mean that if H_1 was rejected in Graph (a) then a fraction 1/2 of w_1 would be recycled to H_2 so that the weight at H_2 would become $w_2 + (1/2)w_1 = 9/10$ and the remainder $(1/2)w_1 = 1/10$ would go to H_3; the weight at H_4 would remain 0 because there is no arrow going from H_1 to H_4 meaning that $g_{14} = 0$. The rejection of H_1 in Graph (a) would lead to Graph (b) with new transition weights obtained from (7.3.2) as: $g_{23} = 1/3$, $g_{24} = 2/3$, $g_{42} = g_{43} = 1/2$ and $g_{32} = 1$. Similarly, if H_2 was initially rejected in Graph (a), then a fraction 1/2 of w_2 would be recycled to

H_1 so that the weight at H_1 would become $w_1 + (1/2)w_2 = 3/5$ and the remainder $(1/2)w_2 = 2/5$ would go to H_4; the weight at H_3 would remain 0 as there is no arrow going from H_2 to H_3 giving $g_{23} = 0$. The rejection of H_2 in Graph (a) would lead to Graph (c) with transition weights obtained from (7.3.2) as: $g_{13} = 2/3$, $g_{14} = 1/3$, $g_{34} = g_{31} = 1/2$ and $g_{41} = 1$.

The value $g_{32} = 1$ in this Graph (b) indicates that if H_3 was rejected after the rejection of H_1 then the entire weight of $(1/2)w_1 = 1/10$ at H_3 would be recycled to H_2, so that the total weight at H_2 after the rejection of both H_1 and H_3 would be $(w_2+(1/2)w_1 = 9/10)+((1/2)w_1 = 1/10) = 1$; the weight at H_4 would remain zero as in this graph there is no arrow going from H_3 to H_4. Therefore, after the rejection of both H_1 and H_3, the Graph (b) would reduce to Graph (d). Similarly, $g_{41} = 1$ in Graph (c) indicates that if H_4 was rejected after the rejection of H_2 then the entire weight $(1/2)w_2 = 2/5$ at H_4 would be recycled to H_1, so that the total weight at H_1 after the rejection of both H_2 and H_4 would be $(w_1 + (1/2)w_2)) + ((1/2)w_2) = 1$; the weight at H_3 would remain zero. Therefore, after the rejection of both H_2 and H_4, the Graph (c) would reduce to Graph (e). However, if either H_2 was rejected in Graph (b) or H_1 was rejected in Graph (c), then these graphs would reduce to Graph (f).

7.3.5 *Illustrative Example of the Graphical Approach for GS Trials*

The above graphical approach originally developed for testing multiple hypotheses of non-GS trials also applies to GS trials. Recycling of alpha of a rejected hypothesis to other hypotheses occurs similarly, but boundary values for testing the unrejected hypotheses are calculated using spending functions. For example, consider the above trial for testing two primary hypotheses H_1 and H_2 and two secondary hypotheses H_3 and H_4, where pairs (H_1, H_3) and (H_2, H_4) are parent–descendant.

In the beginning, we start with Graph (a) of Fig. 7.2 with four hypotheses $\{H_j, j \in \boldsymbol{I}_1 = \{1, 2, 3, 4\}\}$ identified by four nodes and the associated weights $\{w_j(\boldsymbol{I}_1), j \in \boldsymbol{I}_1\} = \{1/5, 4/5, 0, 0\}$. These weights give the starting overall significance levels $\{w_j(\boldsymbol{I}_1)\alpha, j \in \boldsymbol{I}_1; \alpha = 0.025\} = \{0.005, 0.02, 0, 0\}$, and the j-th one for testing of H_j by using its spending function f_j for determining its boundary values for testing. That is, in the beginning, with Graph (a), we test each H_j $(j \in \boldsymbol{I}_1)$ in the univariate GS testing framework for the control of the overall Type I error rate at level $w_j(\boldsymbol{I}_1)\alpha$ so that the total overall Type I error rate control for the trial is at level $\sum_{j\in \boldsymbol{I}_1} w_j(\boldsymbol{I}_1)\alpha = \alpha$.

For this example, we assume that f_j's are all equal to $f(\gamma, t) = 2\{1 - \Phi(z_{1-\gamma/2}/\sqrt{t})\}$, which is OF-like, and γ is the overall significance level for the repeated testing of a hypothesis. The weights $w_3(\boldsymbol{I}_1) = w_4(\boldsymbol{I}_1) = 0$ indicate that H_3 and H_4 are not tested in Graph (a); if they were tested, they would remain unrejected. The following describes how the procedure performs tests of these hypotheses at

Table 7.3 Tests information at the first look at $t_1 = 1/2$ according to Graph (a)

Overall trial α	0.025			
$j \in \boldsymbol{I}_1$	1	2	3	4
$w_j(\boldsymbol{I}_1)$	1/5	4/5	0	0
$w_j(\boldsymbol{I}_1)\alpha$	0.005	0.02	0	0
$\alpha_j\big(w_j(\boldsymbol{I}_1)\alpha, t_1\big)$	0.00007	0.0010	0	0

Note As the p-values $\{p_j(t_1), j \in \boldsymbol{I}_1\}$ exceed their corresponding boundary values, there is no rejection of a hypothesis at this look

different looks and how it recycles the unused alpha of a rejected hypothesis to other unrejected hypotheses.

Tests at the First Interim Look:
Suppose that at the first look, the information fraction is $t_1 = 1/2$. For this example, we assume that the information fraction at a look remains the same for different hypotheses. If this is not the case, the procedure will proceed as discussed in Sect. 7.3.3. The univariate group sequential procedure for testing a hypothesis in a single-hypothesis trial calculates the boundary values for interim looks given the overall significance level α. However, in our case, there are more than one significance levels as $\big\{w_j(I)\alpha, j \in \boldsymbol{I}_1; \alpha = 0.025\big\} = \{0.005, 0.02, 0, 0\}$ assigned to $\{H_j, j \in \boldsymbol{I}_1\}$. These overall significance levels, and the use of the OF-like spending function at $t_1 = 0.5$, then give the boundary values $\{\alpha_j\big(w_j(\boldsymbol{I}_1)\alpha, t_1\big), j \in \boldsymbol{I}_1\} = \{0.00007, 0.0010, 0, 0\}$ for testing $\{H_j, j \in \boldsymbol{I}_1\}$ at the first look. Note that the subscript of t identifies the look number and the subscript j for the hypothesis H_j being tested. Also note that the boundary value of $\alpha_j\big(w_j(\boldsymbol{I}_1)\alpha, t_k\big)$ is a function of the overall significance level $w_j(\boldsymbol{I}_1)\alpha$ assigned to H_j and the information fraction t_k at look k; here $k = 1$.

Suppose that at the first look, the unadjusted p-values $\{p_j(t_1), j \in \boldsymbol{I}_1\}$ associate with $\{H_j, j \in \boldsymbol{I}_1\}$ are such that $p_j(t_1) \geq \alpha_j\big(w_j(\boldsymbol{I}_1)\alpha, t_1\big)$for $j \in \boldsymbol{I}_1$; consequently, the trial will continue to the second look without rejection of a hypothesis at the first look. For recording purposes, one can summarize the above testing information at the first look as in Table 7.3.

Tests at the Second Look:

Suppose that the trial conducts the second look when the information fraction is $t_2 = 3/4$. Since none of the hypotheses was rejected at the first look, we begin with Graph (a) at the second look, by using the same overall significance levels of $\{w_j(\boldsymbol{I}_1)\alpha, j \in \boldsymbol{I}_1\}) = \{0.005, 0.02, 0, 0\}$ that were used at the first look. However, as $t_2 = 3/4$ at the second look, the use OF-like spending function leads to the boundary values of $\{\alpha_j\big(w_j(\boldsymbol{I}_1)\alpha, t_2\big), j \in \boldsymbol{I}_1\} = \{0.00117, 0.0069, 0, 0\}$ for testing H_j for $j \in \boldsymbol{I}_1$. The boundary values for testing H_3 and H_4 remain zero, as there is no rejection of a primary hypothesis so far. Suppose that at this second look, the observed p-values associated with for H_1, H_3, H_2, and H_4 are $p_1(t_2) = 0.001$, $p_2(t_2) = 0.020$, $p_3(t_2) = 0.040$, and $p_4(t_2) = 0.091$, respectively. These results lead to the rejection

Table 7.4 **a** Tests information at the second look at $t_2 = 3/4$ according to Graph (a) after no rejection at the first look. **b** Tests information at the second look at $t_2 = 3/4$ according to Graph (b) after the rejection of H_1 at this look

Overall trial α	0.025 (Table 7.4a)			
$j \in \boldsymbol{I}_1$	1	2	3	4
$w_j(\boldsymbol{I}_1)$	1/5	4/5	0	0
$w_j(\boldsymbol{I}_1)\alpha$	0.005	0.02	0	0
$\alpha_j\big(w_j(\boldsymbol{I}_1)\alpha, t_2\big)$	0.00117	0.0069	0	0
p-values: $p_j(t_2)$	0.001	0.020	0.040	0.091
Overall trial α	0.025 (Table 7.4b)			
$j \in \boldsymbol{I}_2$	–	2	3	4
$w_j(\boldsymbol{I}_2)$	–	9/10	1/10	0
$w_j(\boldsymbol{I}_2)\alpha$	–	0.0225	0.00255	0
$\alpha_j\big(w_j(\boldsymbol{I}_2)\alpha, t_2\big)$	–	0.00802	0.00047	0
p-values: $p_j(t_2)$	0.001	0.020	0.040	0.091

Note H_1 is rejected as $p_1(t_2) = 0.001$ is less than its boundary value of 0.00117 (Table 7.4a)
Note As $p_2(t_2) = 0.020 > 0.00802$ and $p_3(t_2) = 0.040 > 0.00047$, there is no additional rejection at the second look (Table 7.4b)

of H_1 at the second look as $p_1(t_2) = 0.001$ is less than its boundary value of 0.00117; see Table 7.4a.

The above rejection of H_1 at the second look then frees its overall significance level of $w_1(\boldsymbol{I}_1)\alpha) = 0.005$ as unused alpha which is recycled to the remaining three hypotheses for their tests according to Graph (b). This revised graph, constructed after the rejection of H_1, allows retesting of the remaining hypotheses $\{H_j, j \in \boldsymbol{I}_2 = \{2, 3, 4\}\}$ at their corresponding overall significance levels of $\big\{w_j(\boldsymbol{I}_2)\alpha, j \in \boldsymbol{I}_2\big\} = \{-, (9/10)\alpha, (1/10)\alpha, (0)\alpha\} = \{-, 0.0225, 0.00255, 0\}$. Note that the overall significance levels for testing H_2, H_3 are now increased creating the possibility of additional rejections of hypotheses at the second look according to Graph (b). The use OF-like spending function with these updated overall significance levels and $t_2 = 3/4$, then produces the boundary values of $\{\alpha_j\big(w_j(\boldsymbol{I}_2)\alpha, t_2\big), j \in \boldsymbol{I}_2\} = \{-, 0.00802, 0.00047, 0\}$ for testing H_j for $j \in \boldsymbol{I}_2$; see Table 7.4b. However, in this table, as $p_2(t_2) = 0.020 > 0.00802$ and $p_3(t_2) = 0.040 > 0.00047$, there is no additional rejections at the second look. Therefore, the trial moves to the next look which is the final look.

Tests at the Final Look:
After the rejection of H_1 at the second look, the tests for the remaining three hypotheses $\{H_j, j \in \boldsymbol{I}_2\}$ at the final look start with the same Graph (b) and the same overall significance levels of $\big\{w_j(\boldsymbol{I}_2)\alpha, j \in \boldsymbol{I}_2\big\} = \{-, (9/10)\alpha, (1/10)\alpha, (0)\alpha\} = \{-, 0.0225, 0.00255, 0\}$ for testing $\{H_j, j \in \boldsymbol{I}_2 = \{2, 3, 4\}\}$. However, as $t_3 = 1$ at this look, the use of the same OF-like spending function produces the boundary val-

Table 7.5 **a** Tests information at the final look at $t_3 = 1$ according to Graph (b) after the rejection of H_1 at the second look. **b** Tests information at the final look at $t_3 = 1$ according to Graph (f) after the rejection of H_1 at the second look and the rejection of H_2 at the final look

Overall trial α	0.025 (Table 7.5a)			
$j \in \boldsymbol{I}_2$	–	2	3	4
$w_j(\boldsymbol{I}_2)$	–	9/10	1/10	0
$w_j(\boldsymbol{I}_2)\alpha$	–	0.0225	0.00255	0
$\alpha_j\big(w_j(\boldsymbol{I}_2)\alpha, t_3\big)$	–	0.01988	0.00234	0
p-values: $p_j(t_3)$	–	0.012	0.008	0.041
Overall trial α	0.025 (Table 7.5b)			
$j \in \boldsymbol{I}_2$	–	2	3	4
$w_j(\boldsymbol{I}_2)$	–	–	2/5	3/5
$w_j(\boldsymbol{I}_2)\alpha$	–	–	0.010	0.015
$\alpha_j\big(w_j(\boldsymbol{I}_2)\alpha, t_3\big)$	–	–	0.00907	0.013440
p-values: $p_j(t_3)$	–	0.012	0.008	0.041

Note As $p_2(t_3) = 0.0120 < 0.01988$ and $p_3(t_2) = 0.008 > 0.00234$, there is a rejection of H_2 at this look (Table 7.5a)
Note As $p_3(t_2) = 0.008 < 0.00907$, H_3 is also rejected at this look (Table 7.5b)

ues of $\{\alpha_j\big(w_j(\boldsymbol{I}_2)\alpha, t_3\big), j \in \boldsymbol{I}_2\} = \{-, 0.01988, 0.00234, 0\}$ for testing of $\{H_j, j \in \boldsymbol{I}_2\}$ at this look. Suppose that at this final look, the observed p-values associated with for H_3, H_2, and H_4 are $p_2(t_3) = 0.012, p_3(t_3) = 0.008$, and $p_4(t_3) = 0.041$, respectively. These results then lead to the rejection of H_2 at the final look as its $p_2(t_2) = 0.012$ is less than its corresponding boundary value of 0.01988; see Table 7.5a.

Now, as H_1 was rejected at the second look and as H_2 is rejected at the final look, the tests of hypotheses H_3 and H_4 at the final look will be at the increased overall significance levels of $\{w_j(\boldsymbol{I}_3)\alpha, j \in \boldsymbol{I}_3 = \{3, 4\}\} = \{(2/5)\alpha, (3/5)\alpha\} = \{0.010, 0.015\}$ according to Graph (f). These with the OF-like spending function give the boundary values of $\{-, -, 0.00907, 0.01344\}$ for testing $\{H_j, j \in \boldsymbol{I}_3\}$, rejecting also H_3 in this final look, as $p_3(t_2) = 0.008$ is less than 0.00907; see Table 7.5b. Consequently, the remaining H_4 can be tested at this look the at the full overall significance level of $\alpha = 0.025$ which gives the boundary value of 0.0220 for its testing. Therefore, as $p_4(t_2) = 0.041 > 0.0220$ for H_4, the trial stops without the rejection of this hypothesis.

7.4 Testing a Secondary Hypothesis When the Trial Stops After the Rejection of a Primary Hypothesis

Consider, for example, a trial with two looks for testing a primary hypothesis H_1 and a secondary hypothesis H_2 with one interim look and a final look at information fractions t_1 and $t_2 = 1 (0 < t_1 < t_2)$, respectively. The trial, if it rejects H_1 at the interim look, stops at that look for ethical reasons. This will in general be the case when H_1 is associated with an endpoint such as mortality. Therefore, H_2 must be tested at the same interim look when H_1 is rejected, and this test for H_2 must occur after the rejection of H_1.

A question often arises: Can the test of H_2 at the interim look, after the rejection of H_1 at that look, be at the full significance level α (e.g., $\alpha = 0.025$)? This question may arise based on the considerations that H_2 is not tested unless H_1 is first rejected and there is no repeated testing of H_2 after the rejection of H_1. Tamhane et al. (2010) (also Xi and Tamhane 2015) showed that the answer of this question is affirmative, only for the special case when the test statistics for testing H_1 and H_2 are independent. However, this can inflate the overall Type I error rate if the test statistics are correlated. They show that with certain distributional assumptions of the test statistics, the exact adjusted significance level for testing H_2 can be found if this correlation is known. However, if this correlation is unknown, then an upper bound of the adjusted significance levels can be set that covers all correlations. The following revisits this work in some detail because of its importance for clinical trial applications.

We assume that the trial is designed to demonstrate superiority of a new treatment to control such that $H_i : \delta_i \leq 0 (i = 1, 2)$, where δ is the treatment difference parameter. Also, X and Y are the test statistics for testing H_1 and H_2, respectively, which become $(X(t_k), Y(t_k))$ at information times $t_k (k = 1, 2)$. Also, following the results of Sect. 7.2, we assume that each pair $(X(t_1), X(t_2))$ and $(Y(t_1), Y(t_2))$ follows a standard bivariate normal distribution with the same correlation of $\sqrt{t_1}$. Further, we assume that each pair $(X(t_1), Y(t_1))$ and $(X(t_2), Y(t_2))$ follows a standard bivariate normal distribution with correlation coefficient of $\rho \geq 0$. Furthermore, we assume that (c_1, c_2) and (d_1, d_2) are boundary values for testing H_1 and H_2, respectively, so that d_1 is used only when H_1 is rejected at the first look; similarly, d_2 is used only when H_1 being retained at the first look is rejected at the final look. The test strategy for this 2-stage design can then be stated as follows:

Step 1:

$$\text{If } X(t_1) \leq c_1 \longrightarrow \text{Go to Step 2}$$

$$\text{If } X(t_1) > c_1 \longrightarrow \text{Reject } H_1 \text{ and test } H_2$$

$$\text{If } Y(t_1) > d_1 \longrightarrow \text{Reject } H_2\text{; else, retain it}$$

(In either case terminate the trial)

Step 2:

$$\text{If } X(t_2) \leq c_2 \longrightarrow \text{Terminate the trial without any rejection}$$

$$\text{If } X(t_2) > c_2 \longrightarrow \text{Reject } H_1 \text{ and test } H_2$$

$$\text{If } Y(t_2) > d_2 \longrightarrow \text{Reject } H_2\text{; else, retain it.}$$

Determining the Boundary Values of the Procedure
Tests for H_1 and H_2 for the above 2-stage design can be carried out by the method based on the closed testing for GS trials as addressed in Sect. 7.3.2. The intersection hypothesis H_{12} would be tested by the weighted Bonferroni tests with weights of $w_1 = 1$ and $w_2 = 0$ associated with the tests of H_1 and H_2, respectively; $w_2 = 0$ for H_2 implies that this weight can increase only after H_1 is rejected. Therefore, for this design, the rejection of H_1 at level α implies the rejection of H_{12} at level α. Consequently, H_2 can be tested at the full significance level α. But as the trial is a GS trial with one interim look, the boundary values c_1 and c_2 for testing H_1 can then be found from the following two equations:

$$\Pr\{X(t_1) > c_1 | H_1\} = f_1(\alpha, t_1(X))$$

and

$$f_1(\alpha, t_1(X)) + \Pr\{X(t_1) \leq c_1 \cap X(t_2) > c_2 | H_1\} = f_1(\alpha, t_2(X) = 1),$$

where $f_1(\alpha, t)$ is the spending function for testing H_1, and $t_1(X)$ and $t_2(X)$ are the information fractions for testing H_1 at the first and final looks, respectively. For example, when $f_1(\alpha, t)$ is OF-like, $\alpha = 0.025$, and $t_1(X) = 0.5$, then $c_1 = 2.95901$ and $c_2 = 1.96869$ on the normal z-scale which translates to $\alpha_1(0.025, t_1(X) = 0.5) = 0.00153$ and $\alpha_2(0.025, t_2(X) = 1) = 0.02449$ on the p-value scale.

Since the significance level α for the test of H_1 after its rejection recycles to test H_2, the boundary values (d_1, d_2) for H_2 need to be calculated also by a GS method but at the same level α. Reason for this is that, though H_2 is tested after the rejection of H_1, the rejection of H_2, similar to that for H_1, can occur either at the first look or at the final look. Thus, if one uses the Pocock (1977) method for calculating the

boundary values for testing H_2, then at $\alpha = 0.025$, $t_1(Y) = 0.5$ and $t_2(Y) = 1$, the value $d = d_1 = d_2 = 2.17828$ (on the z-scale) which is 0.01469 on the p-value scale. However, the test statistics X and Y in many applications will be positively correlated. Therefore, if this correlation is ρ, and remains the same for the two looks, then it is natural to ask a key question: Is it possible to take advantage of this correlation and find $d^* \leq d$ while maintaining the control of the overall Type I error rate at level $\alpha = 0.025$?

The following shows that this is possible. But the extent of the gain depends on the value of ρ. Larger is the value of ρ on the interval $0 \leq \rho \leq 1$, lesser is the gain, and as ρ approaches one, the value of d^* approaches d determined by the Pocock (1977) method.

Determining the Value of d^*

Testing of H_1 and H_2 gives rise to three null hypotheses configurations $H_{12} = H_1 \cap H_2$, $H_1 \cap K_2$, and $K_1 \cap H_2$, where K_1 and K_2 are alternatives to H_1 and H_2, respectively. The overall Type I error rate for testing H_1 and H_2 under the first two configurations is ≤α. That is, tests for H_1 control this error rate at level α regardless of whether H_2 is true or false. Therefore, we need to find $z_y = d^*$ by solving for z_y in the following equation under $K_1 \cap H_2$.

$$\Pr\{X(t_1) > c_1 \cap Y(t_1) > z_y\} + \Pr\{X(t_1) \leq c_1 \cap X(t_2) > c_2 \cap Y(t_2) > z_y\} = \alpha. \tag{7.4.1}$$

Now, Cov $\{X(t_1), X(t_2)\} = \sqrt{t_1}$, Cov $\{X(t_1), Y(t_2)\} = \sqrt{t_1}\,\rho$, and $\text{Cov}\{X(t_1), Y(t_1)\} = \text{Cov}\{X(t_2), Y(t_2)\} = \rho$. Also, $\text{E}\{X(t_1)\} = \theta\sqrt{t_1}$, $E\{X(t_2)\} = \theta$, and $E\{Y(t_i)\} = 0$ for $i = 1, 2$, because of $K_1 \cap H_2$ and θ being the drift parameter for X. Further, one can show that conditional on $X(t_2) = x(t_2)$, the test statistics $X(t_1)$ and $Y(t_2)$ are independently normally distributed as:

$X(t_1)$ is $N\left\{x(t_2)\sqrt{t_1}, 1 - t_1\right\}$ and $Y(t_2)$ is $N\left\{(x(t_2) - \theta)\rho, 1 - \rho^2\right\}$

Therefore, the Eq. (7.4.1) for finding $z_y = d^*$ can be written as:

$$\alpha = 1 - \Phi\left(c_1 - \theta\sqrt{t_1}\right) - \Phi\left(z_y\right) + \Phi_{12}\left(c_1 - \theta\sqrt{t_1}, z_y; \rho\right) + \int_{c_1-\theta}^{\infty} \Phi\left(\frac{c_1 - \theta\sqrt{t_1} - u\sqrt{t_1}}{\sqrt{1 - t_1}}\right) \Phi\left(\frac{-z_y - u\rho}{\sqrt{1 - \rho^2}}\right) \phi(u)du, \tag{7.4.2}$$

where Φ and ϕ are the density and the cumulative distribution functions of the $N(0,1)$ random variable, and Φ_{12} is the cumulative distribution function of the standard bivariate normal distribution with correlation coefficient of ρ.

Therefore, specifying values of ρ, t_1, c_1, and c_2, one can construct a graph $z_y = f(\theta)$ over the interval $\theta > 0$ that satisfy Eq. (7.4.2). Figure 7.3 shows such graphs for different values of ρ when $\alpha = 0.025$ (1-sided), $t_1 = 0.5$, and $c_1 = 2.95901$ and $c_2 = 1.96869$ on using the OF-like α-spending function. Constructing such a graph for a given ρ then gives $d^* = z_y$ where the maximum occurs for that ρ. Such a

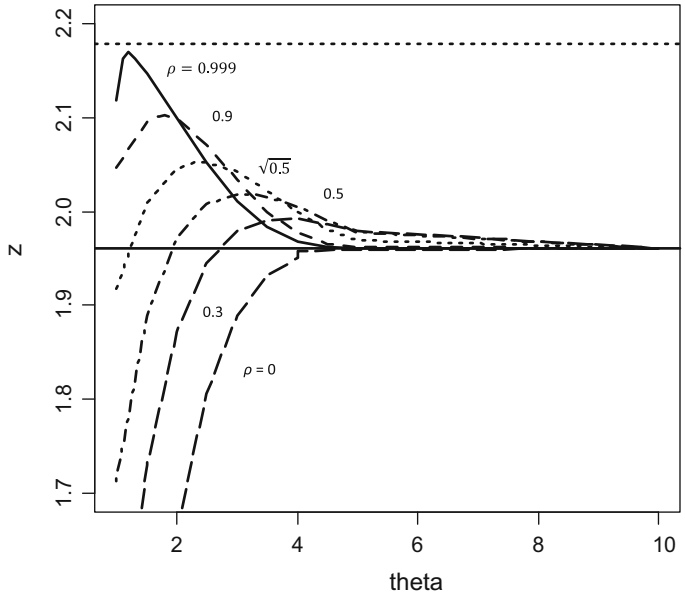


Fig. 7.3 Graph of $z_y = f(\theta)$ over the interval $\theta > 0$ satisfying Eq. (7.4.2). In this graph, theta $= \theta$ and $z = z_y$. The horizontal dashed line in the graph represents the Pocock boundary

selection of d^* assures that the right side of (7.4.2) is $\leq \alpha$ for all $\theta > 0$. Table 7.6, for the above values of α, t_1, c_1, and c_2, gives d^* values and the corresponding α_{d^*} values on the p-value scale for values of ρ shown in column 1 of this table. This table also includes values of θ^* where the d^* values occur. Results of this table show that if the test statistics for testing H_1 and H_2 are uncorrelated, then the test for H_2 at a look after the rejection of H_1 at that look can be at the full significance level α. However, if these test statistics are correlated, then this significance level for testing H_2 is correlation dependent. For positive correlations, this significance level for testing decreases with increasing correlation value and approaches to a value by the Pocock (1977) method.

7.5 Concluding Remarks

Confirmatory clinical trials have been gold standards for establishing efficacy of new treatments. However, such trials when designed with a single primary endpoint do not provide sufficient information when one must assess the effect of the new treatment on different but important multiple characteristics of the disease. For these situations, trials include multiple endpoints related to these disease characteristics and a statistical plan for testing multiple hypotheses on these endpoints for establishing

Table 7.6 Values of d^* for the 2-stage design for different correlations when $\alpha = 0.025$ (1-sided), $t_1 = 0.5$, and $c_1 = 2.95901$ and $c_2 = 1.96869$ on using the OF-like α-spending function

Correlation ρ	d^* (Z-scale)	α_{d^*} (p-value scale)	$\theta = \theta^*$
0.0	1.95996	0.02500	$\theta^* = \text{all}\,\theta > 6.5$
0.1	1.96958	0.02444	4.54
0.2	1.98063	0.02382	4.12
0.3	1.99160	0.02321	4.00
0.4	2.00497	0.02248	3.43
0.5	2.01872	0.02176	3.11
0.6	2.03407	0.02097	2.78
$\sqrt{0.5}$	2.05314	0.02003	2.45
0.8	2.07326	0.01907	2.15
0.9	2.10262	0.01775	1.79
0.99	2.15450	0.01560	1.31
0.999	2.17026	0.01499	1.20
PK value	$d = 2.17828$	$\alpha_d = 0.01469$	–
Conservative		$\alpha/2 = 0.0125$	–

Note $\theta = \theta^*$ is the value of θ where z_y is maximum on the graph $z_y = f(\theta)$ over the interval $\theta > 0$ satisfying Eq. (7.4.2)

efficacy findings of new treatments. However, testing multiple hypotheses in a trial can raise multiplicity issues causing inflation of the Type I error rate. Fortunately, many novel new statistical methods, such as gatekeeping and graphical methods, are now available in the literature for addressing all types of multiplicity issues of clinical trials. These novel methods have advanced the role of statistical methods in designing modern clinical trials with multiple endpoints or multiple objectives.

In clinical trials with serious endpoints, such as death, often a new treatment is added to an existing therapy for detecting a relatively small but clinically relevant improvement in the treatment effect beyond what the existing therapy provides. Designing and conducting such and other trials for serious diseases can be complex, as these trials may require thousands of patients to enroll and several years to complete. Ethical and economic reasons may necessitate that these trials be designed with interim looks for finding the effect of the treatment at an earlier time point allowing the possibility of stopping the trial early when it becomes clear that the study treatment has the desired efficacy or it is futile to continue the trial further. Such trials that allow analyses of the accumulated data at interim looks for the possibility of stopping the trial early for efficacy or futility reasons are commonly known as group sequential trials.

Obviously, interim analyses of the data in a group sequential trial amounts to repeated testing of one or more hypotheses and would result in Type I error rate inflation, so multiplicity adjustment would be required for drawing valid inference. As mentioned in this chapter, several approaches have been cited in the literature for

addressing the control of Type I error rate for repeated tests of a single hypothesis related to a single primary endpoint of the trial. However, approaches for addressing the multiplicity issues for testing multiple hypotheses related to multiple endpoints of group sequential trials are less frequent in the literature.

This chapter, in addition to providing a brief review of procedures and citing key references thereof for the repeated testing procedures of a single endpoint hypothesis in groups sequential trials, considers procedures for handling multiplicity issues for repeated testing of multiple endpoint hypotheses of trials. In this regard, we distinguish two cases of multiple endpoints which guide the approach for handling the multiplicity issue. The first case arises when after a hypothesis is rejected at an interim look, the trial can continue to test other hypotheses at subsequent looks for additional claims. A testing approach for this is to use the Bonferroni inequality which requires splitting the significance level either among the endpoints or among the different looks. This approach is now rarely used because of the low power of the tests.

A better approach (discussed in Sect. 7.3.2) is to consider the use of the closed testing with the weighted Bonferroni tests of the intersection hypotheses, when the weights satisfy the consonance property. This approach allows recycling of the significance level of a rejected hypothesis to the other hypotheses, thus increasing the power of the test procedure. However, as discussed, the recycling of the significance level from a rejected hypothesis to other hypotheses occurs through an α-spending function and is not simple as with non-group sequential trials.

The closed testing-based approach can be manageable when testing 2–3 hypotheses, but it may be difficult to set up for testing more than three hypotheses, for example, when testing two primary and two secondary hypotheses in a trial, as selecting weights for the weighted Bonferroni tests that satisfy the consonance property can be complicated. For these advanced cases, a graphical approach is recommended which is easier to plan, to use, and to communicate to non-statisticians. This chapter illustrates the application of these two approaches through illustrative examples, showing details of the derivations of the significance levels.

The second case arises (discussed in Sect. 7.4), for example, for a group sequential trial designed for testing a primary and a secondary endpoint hypotheses, and the trial stops at an interim look for ethical reasons when the primary hypothesis is rejected at that look in favor of the study treatment. The issue then arises as to what would be the significance level for testing the secondary hypothesis at that look, given that that the secondary hypothesis is tested only after the primary one is rejected first. This issue has been investigated in the literature in detail, but we have revisited it for increasing its awareness, as group sequential trials are frequently designed with a single primary hypothesis and multiple secondary hypotheses. A natural way to address this problem is to use the graphical procedure and recycle the significance level of the rejected primary hypothesis to secondary hypotheses using the Pocock-like α-spending function.

Glimm et al. (2010) illustrated that using the Pocock-like group sequential test to the secondary hypotheses has a power advantage over the O'Brien-Fleming boundary. Other approaches that consider correlation information between the test statistics

can also be used for simple cases, for example, for the case of testing a single primary and a single secondary hypothesis.

Power considerations in designing GS trials that tests multiple hypotheses are also important. However, this topic is beyond the scope of this paper. The power issue would generally be like those for testing multiple hypotheses in a non-GS trial.

Acknowledgements The authors are grateful to Drs. Frank Bretz, Dong Xi, and Estelle Russek-Cohen for providing detailed comments on this chapter which helped in improving the readability of the materials presented.

Disclaimer This paper reflects the views of the authors and must not be construed to represent FDA's views or policies.

References

Alosh, M., Bretz, F., & Huque, M. F. (2014). Advanced multiplicity adjustment methods in clinical trials. *Statistics in Medicine, 33*(4), 693–713.

Armitage, P., McPherson, C. K., Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society Series A, 132*, 235–244.

Bretz, F., Maurer, W., Brannath, W., & Posh, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine, 28,* 586–604.

Bretz, F., Maurer, W., & Maca, J. (2014). Graphical approaches to multiple testing. Chapter 14. In W. Young & D. G. Chen (Eds.), *Clinical trial biostatistics and biopharmaceutical applications* (pp. 349–394). Boca Raton: Chapman and Hall/CRC Press.

Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W., & Rohmeyer, K. (2011). Graphical approaches for multiple comparison procedures using weighted Bonferroni Simes or parametric tests. *Biometrical Journal, 53*(6), 894–913.

SAS Online Doc. *Version 8*. Copyright 1999 by SAS Institute Inc., Cary, NC, U.S.A.

Dmitrienko, A., Offen, W. W., & Westfall, P. H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine, 22,* 2387–2400.

Dmitrienko, A., & Tamhane, A. C. (2009). Gatekeeping procedures in clinical trials. In A. Dmitrienko, A. C. Tamhane, & F. Bretz (Eds.), *Multiple testing problems in pharmaceutical statistics* (Chap. 1). Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series.

Dmitrienko, A., Tamhane, A. C., & Wiens, W. (2008). General multi-stage gatekeeping procedures. *Biometrical Journal, 50,* 667–677.

EAST software 6.3. (2014). By Cytel Software Corporation, Cambridge, MA, U.S.A.

Ellenberg, S. S., Fleming, T. R., & DeMets, D. L. (2017). *Data monitoring committees in clinical trials*. New York: Wiley.

Emerson, S. (2007). Frequentist evaluation of group sequential trial designs. *Statistics in Medicine, 26,* 5047–5080.

Follmann, D. A., Proschan, M. A., & Geller, N. L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics, 50,* 325–336.

Food and Drug Administration. (2017). Guidance for industry: Multiple endpoints in clinical trials. Retrieved May 3, 2017 from https://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm536750.pdf.

Glimm, E., Maurer, W., & Bretz, F. (2010). Hierarchical testing of multiple endpoints in group sequential trials. *Statistics in Medicine, 29,* 219–228.

Hellmich, M. (2001). Monitoring clinical trials with multiple arms. *Biometrics, 57,* 892–898.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple significance testing. *Biometrika, 75,* 800–802.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.

Hommel, G., Bretz, F., & Maurer, W. (2007). Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine, 26,* 4063–4073.

Huque, M. F., Dmitrienko, A., & D'Agostino, R. (2013). Multiplicity issues in clinical trials with multiple objectives. *Statistics in Biopharmaceutical Research, 5*(4), 321–337.

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials.* Boca Raton, FL: Chapman and Hall/CRC.

Lan, K. K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika, 70,* 659–663.

Lan, K. K. G., & Wittes, J. (1988). The B-value: A tool for monitoring data. *Biometrics, 44,* 579–585.

Marcus, R., Peritz, E., & Gabriel, K. R. (1976). On closed testing procedure with special reference to ordered analysis of variance. *Biometrika, 63,* 655–660.

Maurer, W., & Bretz, F. (2013). Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research, 5*(4), 311–320.

Maurer, W., Glimm, E., & Bretz, F. (2011). Multiple and repeated testing of primary, co-primary and secondary hypotheses. *Statistics in Biopharmaceutical Research, 3,* 336–352.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 5,* 549–556.

O'Neill, R. T. (1997). Secondary endpoints cannot be validly analyzed if primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials, 18,* 550–556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis clinical trials. *Biometrika, 64,* 191–199.

Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials. A unified approach*. New York: Springer.

Stallard, N., & Friede, T. (2008). A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine, 27,* 6209–6227.

Tamhane, A. C., Mehta, C. R., & Liu, L. (2010). Testing a primary and a secondary endpoint in a group sequential design. *Biometrics, 66,* 1174–1184.

Tang, D. L., & Geller, N. L. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics, 55,* 1188–1192.

Wassmer, G., & Brannath, W. (2016). *Group sequential and confirmatory adaptive designs in clinical trials*. Switzerland: Springer International Publishing AG.

Whitehead, J. (1997). *The design and analysis of sequential clinical trials*. Chichester: Wiley.

Xi, D., Glimm, E., Bretz, F. (2016). *Multiplicity in chapter 3 of cancer clinical trials: Current and controversial issues in design and analysis*. In: S. L. George, X. Wang, & H. Pang. CRC Press, Taylor and Francis Group.

Xi, D., & Tamhane, A. C. (2015). Allocating recycled significance levels in group sequential procedures for multiple endpoints. *Biometrical Journal, 57*(1), 90–107.

Ye, Y., Li, A., Liu, L., & Yao, B. (2013). A group sequential Holm procedure with multiple primary endpoints. *Statistics in Medicine, 32,* 1112–1124.

Zhu, L., Ni, L., & Yao, B. (2011). Group sequential methods and software applications. *The American Statistician, 65*(2), 127–135.

Chapter 8
Expanded Statistical Decision Rules for Interpretations of Results of Rodent Carcinogenicity Studies of Pharmaceuticals

Karl K. Lin and Mohammad A. Rahman

8.1 Introduction

The Center for Drug Evaluation and Research of the U.S. Food and Drug Administration (CDER/FDA) draft Guidance for Industry: Statistical Aspects of the Design, Analysis, and Interpretation of Chronic Rodent Carcinogenicity Studies of Pharmaceuticals (U.S. department of Health and Human Services, 2001) was announced in Federal Register (Tuesday, May 8, 2001, Vol. 66, No. 89) for a 90-day public comment in 2001. Comments on the document from 16 drug companies, professional organizations of the pharmaceutical industry, and individual experts from USA, Europe, and Japan were submitted to the Agency. The public comments are in Food and Drug Administration Docket No. 01D–0194 and are available to FDA scientists to review. The public comments were positive on the contents and the usefulness of the guidance. The statistical methods recommended in the draft guidance for industry document have been closely followed by statistical reviewers within CDER/FDA and statisticians in drug companies in USA and abroad in their data analyses of carcinogenicity studies.

It is noted that the 2001 draft guidance for industry document discussed the recommended statistical methods for the design, analysis, and interpretation of results

The article reflects the views of the authors and should not be construed to represent FDA's views or policies. Parts of this book chapter were also included in a manuscript that have been published online in the Journal of Biopharmaceutical Statistics.

K. K. Lin (✉) · M. A. Rahman
Division of Biometrics 6, Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Room 4677, Building 21, 10903 New Hampshire Avenue, Silver Spring, MD 20993-0002, USA
e-mail: karl.lin@fda.hhs.gov

M. A. Rahman
e-mail: Mohammad.Rahman@fda.hhs.gov

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_8

only for NDA and IND submissions that include two chronic (long-term) rodent carcinogenicity studies of pharmaceuticals in rats and mice. Two more types of NDA and IND submissions other than those with two chronic studies in rats and mice were either not discussed or barely mentioned in the guidance document. They are submissions including one chronic study in rats and one six-month study in transgenic mice and submissions including only one chronic study in either rats or mice.

The general statistical methods recommended for the analysis of data of chronic carcinogenicity studies of pharmaceuticals in the original draft guidance for industry document are still fairly up-to-date and applicable in general to those other types of submissions. However, it is noted that the 2001 draft guidance for industry document discussed the recommended statistical methods for the interpretation of study results only for NDA and IND submissions that include two chronic rodent carcinogenicity studies of pharmaceuticals in rats and mice. It has become necessary to update the part of methods of interpretation of study results of the original guidance document to include recently recommended interpretation methods for the two other types of NDA and IND submissions.

Furthermore, it has also become necessary to update the part of the recommended interpretation methods in the original guidance document to reflect the more restricted decision rules (explained below) having been used by some practicing pharmacologists/toxicologists in their final determination of the carcinogenic potential of a new drug after the issuance of the 2001 draft guidance document for public comment.

It is specifically recommended throughout the draft guidance document that the trend tests (testing the null hypothesis of no positive trend versus the alternative hypothesis of a positive trend in incidence rates), which have been studied extensively in the literature and internally within CDER/FDA (Lin 1988, 1995, 2000a, b, 2010; Lin and Ali 2006; Lin and Rahman 1998; Lin et al. 2010, 2016; Rahman and Lin 2008, 2009, 2010), should be the primary tests for the evaluation of carcinogenic effects of a drug because they have more statistical power to detect a true carcinogenic effect and that the pairwise comparison tests (testing the null hypothesis of no drug-induced increase versus the alternative hypothesis of a drug-induced increase in incidence rates of a treated group over the control group) should be used only in rare situations in which a comparison between the control group and an individual treated group is considered as more appropriate than the trend test in the carcinogenicity evaluation of a drug. Studying the special nature and the designs of carcinogenicity studies and the need to balance the Type I and Type II errors in statistical inference in a regulatory environment, statistical decision rules were developed by statisticians within the Office of Biostatistics (OB) in CDER/FDA for applying trend tests alone and for applying pairwise comparison tests alone in a statistical review and evaluation of carcinogenicity studies of a new drug.

The draft guidance recommended in the section of interpretation of study results that a positive dose–response (or trend) in incidence rates of an individual tumor type be tested at 0.005 or 0.025 significance level and that a pairwise comparison of an increase in incidence rates in a treated group over the control group of the tumor type be tested at 0.01 or 0.05 significance level for a common or a rare tumor type, respectively, in a standard submission of two chronic studies in both sexes of

rats and mice. A tumor type is defined as common if it has a spontaneous rate of greater than or equal to 1% and as rare otherwise. The use of these decision rules for individual tumor types results in an overall false-positive rate of about 10% in a standard submission containing two chronic studies in both sexes of rats and mice (Lin and Rahman 1998; Rahman and Lin 2008). The 10% overall false-positive rate has been determined as the most appropriate in the regulatory review environment with the consideration of the nature of the designs of carcinogenicity studies of new drugs.

In order to further reduce the false-positive rate (measuring the producer's risk in regulatory reviews of toxicology studies), some practicing pharmacologists/toxicologists have used a so-called joint test in their determination if a test of a drug effect on the development of an individual tumor type is statistically significant. In the joint test, the results of both the test for the positive trend and the pairwise comparison test for an increase in the high dose group over the control group in incidence rates of the individual tumor type have to be statistically significant simultaneously at the levels of significance recommended for the separate tests in the 2001 draft guidance for industry document in order to consider the drug effect on the development of the individual tumor type as statistically significant. We have concerns that the use of levels of significance recommended in the guidance document is for the trend tests alone, and for the pairwise comparison test alone, and not for the joint test; and that there is a serious consequence of huge inflations of the false-negative rate (measuring the consumer's risk in regulatory reviews of toxicology studies) in the use of the joint test with the levels of significance recommended for the separate tests in the 2001 draft guidance document in the final interpretation of the results of a carcinogenicity study.

We have done extensive simulation studies to investigate the impacts of the use of the levels of significance recommended in the 2001 draft guidance document for the separate tests in the joint test on the determination of carcinogenic potential of a new drug and to recommend a new set of decision rules (levels of significance) that balance the false-positive and the false-negative rates at acceptable levels for the use of the joint test. Therefore, it has also become necessary to update the part of recommended methods of interpretation of study results in the 2001 draft guidance document to include those newly recommended sets of decision rules for the joint tests in different types of new drug application submissions.

This book chapter includes two major parts. The first part, serving as the important statistical basis for the second part, includes presentations of results of our simulation studies investigating the impacts of the joint test using the levels of significance recommended in the 2001 draft guidance document for the separate tests on the determination of carcinogenic potential of a new drug. The second part includes our recommended sets of expanded decision rules for the separate tests and the joint tests to be used in those three types of new drug application submissions. Methods of our simulation studies comparing false-negative rates resulting from the trend test alone and from the trend test and the control-high group pairwise comparison test simultaneously (i.e., the joint test) in the determination of the carcinogenicity of new drugs are presented in Sect. 8.2. Results of the simulation studies are presented in

Sect. 8.3. Expanded sets of new decision rules recommended for the separate and the joint tests for the three different types of submission are presented in Sect. 8.4.

8.2 Methods of Our Simulation Studies Comparing False-Negative Rates Resulting from the Trend Test Alone and from the Trend Test Jointly with Control-High Group Pairwise Comparison Test in the Determination of the Carcinogenicity of New Drugs

Two simulation studies were conducted. The objective of the simulation studies is to evaluate the extent of inflation of the false-negative rates resulting from the use of the joint test with the levels of significance recommended for the separate tests in the 2001 draft guidance document for determining a statistically significant drug effect on the development of a given tumor type. The two simulation studies used survival and tumor data generated from the Weibull distribution and the Gamma distribution, respectively, to evaluate the false-negative rates resulting from the use of the following three types of decision rules for determining if a test of the drug effect on development of a given tumor type is statistically significant at the significance levels for separate tests recommended in the 2001 draft guidance for industry document: (a) requiring a statistically significant result in the trend test alone (this is the rule recommended in the draft guidance for industry document); (b) requiring statistically significant results both in the trend test and in any of the three pairwise comparison tests (control versus low, control versus medium, and control versus high); and (c) requiring statistically significant results both in the trend test and in the control versus high group pairwise comparison test (this is the joint test rule). The results of decision rule type (b) were also included the tabulations, but they are not included in the evaluations of the inflations of false-negative rates in this book chapter.

8.2.1 Simulation Study Based on the Weibull Distribution

The first simulation study used the same Weibull distribution and the same sets of values for the parameters used in the National Toxicology Program (NTP) study (Dinse 1985) to reflect various simulated conditions on the effect of early or late tumor appearance, the effect of spontaneous rate, the effect of dose on mortality, and the effect of dose on tumor prevalence. Also in this and next simulations studies, the same NTP assumption that tumor types considered are occult and incidental (nonfatal) was also used. Because of the assumption of the data, the death-rate (life-table) method (Peto et al. 1980) for fatal tumors was not used, and therefore, the censoring process was not considered.

In our first and the NTP studies, the tumor detection time (T_0 in weeks) and the time to natural death (T_1 in weeks) of an animal were modeled by four parameter Weibull distributions as

$$S_l(t|x) = P(T_l > t|X = x) = e^{\{-(C_l + D_l x)[(t - A_l)]^{B_l}\}} \text{ if } t > A_l, \text{ and } S_l(t|x) = 1 \text{ if } t \le A_l$$

where A_l is the location parameter, B_l is the shape parameter, C_l is the baseline scale parameter, D_l is the dose-effect parameter, and $l = 0, 1$. Table 8.1 lists the values of the parameters that were used in Dinse (1985) and in this simulation study. Notes of Table 8.1 also show the factors used in the simulation study.

The prevalence function for incidental tumors equals the cumulative function of for time to tumor onset, i.e.,

$$P(t, x) = 1 - S_0(t|x)$$

The factors used in the NTP study are defined as follows: (1) low or high tumor background rate: The prevalence rate at 2 years in the control group is 5% (low) or 20% (high). (2) Tumors appear early or late: The prevalence rate of the control group at 1.5 years is 50% (appearing early) or 10% (appearing late) of the prevalence rate at 2 years. (3) None, small, or large effect on tumor prevalence: The prevalence of the high dose group at 2 years minus the prevalence of the control group at 2 years is 0% (none effect), or 10% (small effect), or 20% (large effect). (4) None, small, or large effect on mortality: The expected portion of animals alive in the high dose group at 2 years is 70% (none), 40% (small effect), or 10% (large effect). The expected portion of animals alive in the control group at 2 years is taken as 70%.

There are important differences in study design between the NTP study and our study. The NTP study simulated three treatment groups with doses $x = 0$, $x = 1$, and $x = 2$ (called the control, low, and high dose groups), and our study used four treatment groups (with doses $x = 0$, $x = 1$, $x = 2$, and $x = 3$, called the control, low, mid, and high dose groups, respectively). Since the values of the parameters A, B, C, and D used were the same in the two studies, the characterizations of the effects of the dose level on tumor prevalence, factor 3, and on mortality, factor 4, apply to the dose level $x = 2$, i.e., to the mid-dose level in our study. To recast these descriptions in terms of the effect at the $x = 3$ (high dose) level, factors (3) and (4) become factors (3') and (4') described below:

(3'). No dose effect, a small dose effect, or a large dose effect on tumor prevalence: The prevalence of the high dose group ($x = 3$) at 2 years minus the prevalence of the control group at 2 years is 0% (no effect), or approximately 15% (small effect), or approximately 28% (large effect).

(4'). No dose effect, a small dose effect, or a large dose effect on mortality: The expected proportion of animals alive in the high dose group at 2 years is 70% (no effect), 30% (small effect), or 4% (large effect). The expected proportion of animals alive in the control group at 2 years is taken as 70%.

Table 8.1 Values of parameters for the Weibull models (Taken from Dinse 1985) in the generation of the simulated survival and tumor data

Weibull parameters				
Time to death (T_1 in weeks)				
Drug effect on death	A	B	Scale $\times 10^4$	
			C	D
None	0	4	0.0000305	0
Small	0	4	0.0000305	0.0000239
Large	0	4	0.0000305	0.00008325

Time to tumor onset (T_0 in weeks) corresponding to each model for time to death						
Background tumor rate	Tumor appearance	Dose effect on tumor prevalence	A	B	Scale x 10^4	
					C	D
Low	Early	None	17	2	0.0678	0
Low	Early	Small	17	2	0.0678	0.0736
Low	Early	Large	17	2	0.0678	0.1561
Low	Late	None	56	3	0.00465	0
Low	Late	Small	56	3	0.00465	0.005025
Low	Late	Large	56	3	0.00465	0.010675
High	Early	None	21	2	0.324	0
High	Early	Small	21	2	0.324	0.097
High	Early	Large	21	2	0.324	0.209
High	Late	None	57	3	0.0215	0
High	Late	Small	57	3	0.0215	0.00645
High	Late	Large	57	3	0.0215	0.01383

Notes on factors used in the simulation: (1) low or high tumor background rate: The prevalence rate at 2 years in the control group is 5% (low) or 20% (high). (2) Tumors appear early or late: The prevalence rate of the control group at 1.5 years is 50% (appearing early) or 10% (appearing late) of the prevalence rate at 2 years. (3) None, small, or large effect on tumor prevalence: The prevalence of the high dose group at 2 years minus the prevalence of the control group at 2 years is 0% (none effect), or 10% (small effect), or 20% (large effect). (4) None, small, or large effect on mortality: The expected portion of animals alive in the high dose group at 2 years is 70% (none), 40% (small effect), or 10% (large effect). The expected portion of animals alive in the control group at 2 years is taken as 70%

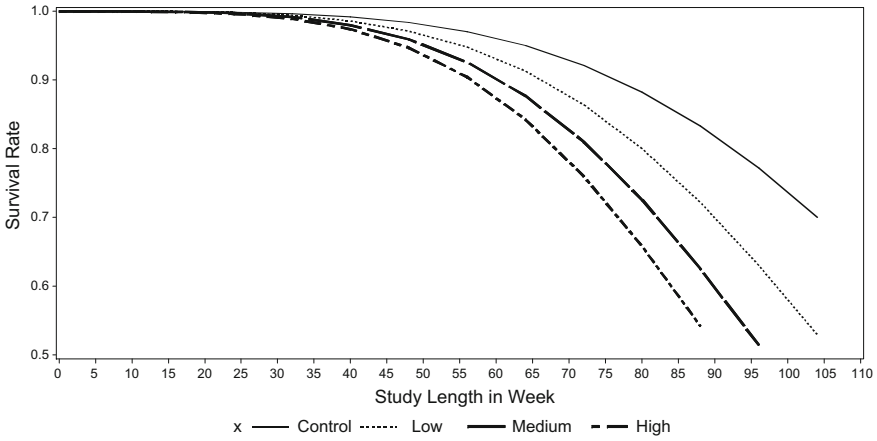


Fig. 8.1 Weibull model used to generate the survival data when the dose effect on mortality is small

These differences in design between the NTP and our studies can be expected to have the following effects on the Type 2 error rates for this and the other studies of ours (relative to the NTP study):

The higher tumor prevalence rates in the high dose groups should help to reduce the false-negative rates (or to increase the levels of power) of statistical tests.

On the other hand, higher levels of mortality will reduce the effective sample size and thus tend to increase the false-negative rates (or to decrease the levels of power).

Two vectors, each with 200×1 dimensions, T_0 and T_1, were generated from two Weibull distributions representing tumor onset time and the time to death of 200 animals. The actual time of death (T) for the animal was defined as the minimum of T_1 and 104 weeks, i.e., $T = (\min(T_1, 104))$. The animal developed the tumor ($Y = 1$) if the time to tumor onset did not exceed the time to death ($T_0 \leq T$), and ($Y = 0$) otherwise. The actual tumor detection time was assumed to be the time of death. Animals in the same dose group were assumed to be equally likely to develop the tumor in their lifetimes. It was assumed that tumors were developed independently of each other. Figure 8.1 contains the graphical presentations of the Weibull models used to generate the survival data when the dose effect on mortality is small, and Fig. 8.2 contains the graphical presentations of the Weibull models used to generate the tumor prevalence data when the background tumor rate is low, the dose effect on tumor prevalence is large, and the tumor appears early in our study.

The age-adjusted prevalence method (Peto et al. 1980) for dose–response relationship and age-adjusted Fisher exact test for pairwise comparisons in tumor incidence using the NTP partition of time intervals were applied to the generated survival and tumor incidence dataset of the 200 animals. The p values of the trend test and of the pairwise comparisons between the control and each of the individual treated groups were recorded and compared with the recommended levels of significance recommended in the 2001 draft guidance for industry document for the trend test alone

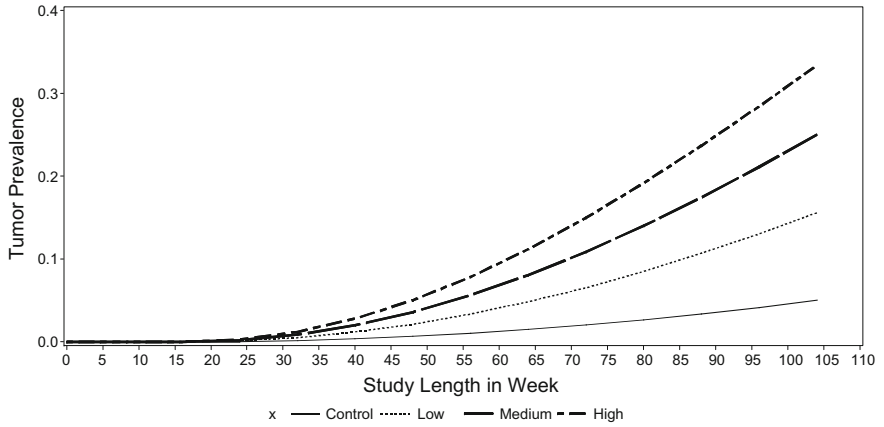


Fig. 8.2 Weibull model used to generate the tumor prevalence data when the background tumor rate is low, the dose effect on tumor prevalence is large, and the tumor appears early

and pairwise comparisons alone for the three decision rules in determining if a drug effect on the development of a given tumor type is statistically significant.

The levels of significance are those with adjustment for effect of multiple testing for the test for positive trend recommended in Lin and Rahman (1998), U.S. Department of Health and Human Services (2001), and Rahman and Lin (2008), and those for pairwise comparisons suggested by Haseman (1983, 1984a, b). False-positive (Type I) and false-negative (Type II) rates were estimated after 10,000 simulation runs for each of the 36 simulation conditions.

8.2.2 Simulation Study Based on the Gamma Distribution

The second simulation study was based on the Gamma distribution under the same simulation conditions as those under the first simulation study based on the Weibull distribution. The purpose of the second simulation study is to compare the results of this study with those from the study based on the Weibull distribution to improve the validity of the simulation results of the first simulation study. For a valid comparison of results, the experimental conditions, the methods of analysis, and the decision rules used in this second simulation study are similar to those used in the first simulation study based on the Weibull distribution. The values of the parameters of the Gamma distribution used in this simulation study are presented in Table 8.2.

Table 8.2 Values of parameters for the gamma distribution in the generation of the simulated survival and tumor data
$y = \frac{\beta^{(\alpha-\delta^* x)}}{\Gamma(\alpha-\delta^* x)} e^{-\beta t} t^{(\alpha-\delta^* x)-1}$ where x is the dose level and $t > \infty$

Gamma parameters for survival data generation				
Background survival rates	Dose effects on death	α	β	δ
70%	None	18	0.148	0
	Small	18	0.148	1.01
	Large	18	0.148	2.22

Gamma parameters for tumor incidence data generation					
Background survival rates	Tumor appearance	Dose effects on death	α	β	δ
Low (5%)	Early	None	3	0.00800	0
		Small	3	0.00800	0.30
		Large	3	0.00800	0.45
	Late	None	16	0.10028	0
		Small	16	0.10028	0.80
		Large	16	0.10028	1.20
High (20%)	Early	None	3	0.01445	0
		Small	3	0.01445	0.22
		Large	3	0.01445	0.35
	Late	None	18	0.13500	0
		Small	18	0.13500	0.60
		Large	18	0.13500	1.08

8.3 Results of Our Simulation Studies

8.3.1 Results of Our First Simulation Study Based on the Weibull Distribution

Table 8.3 contains the estimated (attained) false-negative rates resulting from our extensive simulation using the Weibull distribution under the following requirements for concluding a test result as a statistically significant effect: (1) requiring a statistically significant result in the trend test alone, (2) requiring statistically significant results in the trend test and in any of the C-L, C-M, and C-H pairwise comparison tests simultaneously, and (3) requiring statistically significant results in the trend test and in the C-H pairwise comparison tests simultaneously. The last two columns of the table show the percent changes of the error rate of (2) over that of (1), and of the error rate of (3) over that of (1), respectively.

It is noted that the rates in the table obtained from simulation conditions 1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, and 34 are actually one minus the estimated false-positive

Table 8.3 Estimated false negative rates of trend test alone and trend test along with pairwise comparisons in simulations based on Weibull distribution

Dose				Back	False-negative rate				
Simulation condition number	Effect on death	Tumor appearance	Dose effect on tumor rate	Ground tumor rate	Trend	Trend and high	Trend and any	Percent change Tr-high	Percent change Tr-any
1	No	Early	No	0.0500	0.9840	0.9934	0.9919	0.9553	0.8028
2	No	Early	Small	0.0500	0.6283	0.7084	0.6957	12.75	10.73
3	No	Early	Large	0.0500	0.1313	0.1780	0.1595	35.57	21.48
4	No	Late	No	0.0500	0.9827	0.9927	0.9915	1.018	0.8955
5	No	Late	Small	0.0500	0.6314	0.7208	0.7076	14.16	12.07
6	No	Late	Large	0.0500	0.1408	0.2018	0.1811	43.32	28.62
7	No	Early	No	0.2000	0.9953	0.9979	0.9974	0.2612	0.2110
8	No	Early	Small	0.2000	0.8377	0.8805	0.8715	5.109	4.035
9	No	Early	Large	0.2000	0.3424	0.4270	0.3980	24.71	16.24
10	No	Late	No	0.2000	0.9952	0.9972	0.9972	0.2010	0.2010
11	No	Late	Small	0.2000	0.8399	0.8869	0.8772	5.596	4.441
12	No	Late	Large	0.2000	0.3754	0.4864	0.4565	29.57	21.60
13	Small	Early	No	0.0500	0.9855	0.9985	0.9978	1.319	1.248
14	Small	Early	Small	0.0500	0.6967	0.8465	0.8324	21.50	19.48
15	Small	Early	Large	0.0500	0.2152	0.4112	0.3574	91.08	66.08
16	Small	Late	No	0.0500	0.9819	0.9991	0.9977	1.752	1.609
17	Small	Late	Small	0.0500	0.7220	0.9161	0.8903	26.88	23.31

(continued)

Table 8.3 (continued)

Dose				Back	False-negative rate				
Simulation condition number	Effect on death	Tumor appearance	Dose effect on tumor rate	Ground tumor rate	Trend	Trend and high	Trend and any	Percent change Tr-high	Percent change Tr-any
18	Small	Late	Large	0.0500	0.2682	0.6794	0.6021	153.3	124.5
19	Small	Early	No	0.2000	0.9948	0.9996	0.9995	0.4825	0.4725
20	Small	Early	Small	0.2000	0.8753	0.9694	0.9606	10.75	9.745
21	Small	Early	Large	0.2000	0.4649	0.7564	0.7110	62.70	52.94
22	Small	Late	No	0.2000	0.9961	0.9999	0.9996	0.3815	0.3514
23	Small	Late	Small	0.2000	0.8935	0.9939	0.9885	11.24	10.63
24	Small	Late	Large	0.2000	0.5380	0.9455	0.9095	75.74	69.05
25	Large	Early	No	0.0500	0.9856	0.9994	0.9989	1.400	1.349
26	Large	Early	Small	0.0500	0.8381	0.9587	0.9480	14.39	13.11
27	Large	Early	Large	0.0500	0.5358	0.8133	0.7796	51.79	45.50
28	Large	Late	No	0.0500	0.9828	1.000	1.000	1.750	1.750
29	Large	Late	Small	0.0500	0.8675	0.9960	0.9886	14.81	13.96
30	Large	Late	Large	0.0500	0.6447	0.9807	0.9428	52.12	46.24
31	Large	Early	No	0.2000	0.9940	1.000	1.000	0.6036	0.6036
32	Large	Early	Small	0.2000	0.9414	0.9994	0.9985	6.161	6.065

(continued)

Table 8.3 (continued)

Dose				Back	False-negative rate				
Simulation condition number	Effect on death	Tumor appearance	Dose effect on tumor rate	Ground tumor rate	Trend	Trend and high	Trend and any	Percent change Tr-high	Percent change Tr-any
33	Large	Early	Large	0.2000	0.7445	0.9823	0.9700	31.94	30.29
34	Large	Late	No	0.2000	0.9956	1.000	1.000	0.4419	0.4419
35	Large	Late	Small	0.2000	0.9585	1.000	0.9999	4.330	4.319
36	Large	Late	Large	0.2000	0.8350	0.9998	0.9989	19.74	19.63

The estimated false-negative rates under simulation numbers 1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, and 34 are actually one minus the estimated false-positive rates because the assumption of no dose effect on tumor prevalence rate is used in those simulations. Columns under (a) "Trend," (b) "Trend and High," and (c) "Trend and Any" list the (1—false-positive rates) or the false-negative rates, respectively, from requiring statistically significant results of the trend test alone, of the trend test and C-H pairwise comparison test simultaneously, and of the trend test and any of the three (C-L, C-M, C-H) pairwise comparison tests. The last two columns list the percent changes of false-negative rate of (b) over (a) and (c) over (a), respectively

rates. In those simulation conditions, it was assumed that there was no dose effort on tumor rate (the null hypothesis of no positive dose–response or increase in incidence in a treated group over the control). The estimated false-negative rates are from the other 24 simulation conditions.

Results of the evaluation of Type I error patterns in the study conducted and reported in Dinse (1985) show that the Peto trend test without continuity correction and with the partition of time intervals of the study duration (0–52, 53–78, 79–92, 93–104 weeks, and terminal sacrifice) proposed by NTP yields attained levels of false-positive rate close to the nominal levels (0.05 and 0.01) used in the test.

The results of our first simulation study show a very interesting pattern in levels of attained Type I error. The attained levels of Type I error under various simulation conditions were divided into two groups. The division is made based on the factor of background rate, either 5% or 20%. The levels of the first group were around 0.005, and they were from the simulated conditions with 20% background rate. The attained Type I levels of the second groups were around 0.015, and they were from the simulated conditions with 5% background rate. The results and the pattern of the attained Type I errors in our first simulation study make sense. For the simulated conditions with 20% background rate, almost all of the 10,000 generated datasets (each dataset containing tumor and survival data of four treatment groups of 50 animals each group) will have tumor rate of equal to or greater than 1% (the definition of a common tumor) in the control group. The attained levels of Type I error under various simulated conditions in this group are closed to the nominal level of significance recommended in the 2001 draft guidance document for common tumors after adjusting the effect of multiple testing. The attained Type I errors of the other group were between the nominal levels of significance of 0.005 (for the trend test for common tumors) and 0.025 (for the trend test for rare tumors) and not around 0.005. The reason for this phenomenon is that, though the background rate in the simulated conditions for this group was 5% that is considered as a rate for a common tumor, some of the 10,000 generated datasets will have tumor rates less than 1% in the control group. For this subset of the 10,000 datasets, the nominal level of 0.025 was used in the trend test.

It becomes more complicated in the evaluation of the attained Type I errors in the use of the joint decision rule. It is so because the two tests are not independent since the pairwise comparison tests used a subset (a half) of the data used in the trend test. Theoretically, if the trend test and the pairwise comparison test are actually independent and are tested at 0.005 and 0.01 levels of significance, respectively, then the nominal level of significance of the joint tests should be $0.005 \times 0.01 = 0.00005$. Some of the levels of attained Type I error of the joint tests are larger than 0.00005 due to the dependence of the two tests that were applied simultaneously.

As mentioned previously, the evaluation of the patterns of the Type II error is the main objective of our simulation studies. As was expected, the false-negative rates resulting from the joint test using the levels of significance recommended for the separate tests recommended in the 2001 draft guidance document are higher than those from the procedure recommended in the draft guidance document that requires only a statistically significant result in the trend test alone. This is due to the fact in

statistics that the false-positive rate and the false-negative rate run in the opposite direction and that the use of the joint test with the levels of significance recommended for the separate tests in the 2001 draft guidance document in the joint test to cut down the former rate will definitely inflate the latter rate.

The magnitude of the inflation of false-negative rate resulting from the trend test alone or a pairwise comparison test alone or both the trend test and the pairwise comparison test between the control and high groups simultaneously (i.e., the joint test) depends on all the four factors, namely drug effect on mortality, background tumor rate, time of tumor appearance, and drug effect on tumor incidence considered in the simulation and also listed in the notes in the bottom of Table 8.1.

Results of this simulation study show that the factor of the effect the dose has on tumor prevalence rate has the largest impact on the inflation of the false-negative rate when both the trend test and the pairwise comparison tests between the control and high groups are required to be statistically significant simultaneously in order to conclude that the effect is statistically significant. The inflations are most serious in the situations in which the dose has a large effect on tumor prevalence. The inflation of false-negative rates resulting from the joint test in this first simulation study based on the Weibull distribution can be as high as 153.3% (i.e., more than two and half times) of that when the trend test alone is required to be statistically significant to consider that the effect is statistically significant. As the results of the second simulation based on the Gamma distribution discussed in next subsection show, the inflation of the false-negative rates resulting from the joint test can go much higher than 153.3%.

The above finding is the most alarming result among those from our first simulation study (and also among those from our second simulation study discussed below). When the dose of the test new drug has large effects on tumor prevalence, it is a clear indication that the drug is carcinogenic. Exactly in these most important situations, the use of the joint test with the levels of significance recommended for the separate tests in the 2001 draft guidance document causes the most serious inflation of the false-negative error rate (or the most serious reduction in statistical power to detect the true carcinogenic effect). The net result of this alarming finding is that the use of the joint test with improper levels of significance can be up to more than two and half times likely to fail to detect a true carcinogenic effect than the procedure based on the result of the trend test alone.

It is true that the results in Table 8.3 also show that for the situations in which the dose has a small effect on tumor prevalence, the increases of false-negative rates caused by the joint test using the levels of significance for the separate tests recommended in the 2001 draft guidance document are not much more (be up to 26%) than those from using the trend test alone. However, this observation does not imply that the use of the levels of significance recommended in the 2001 draft guidance document for the separate tests in the joint test is justified. The reason is that the small group sizes are used as a surrogate of a large population with low tumor incidence endpoint in the standard carcinogenicity studies. There is almost no power (or with false-negative rates close to 100%) for a statistical test to detect a true effect because of the small group size and the low tumor incidence rates and the small

dose effect on tumor prevalence. In those situations, there will be little room for the further increase in false-negative rate no matter how many additional requirements of statistical significance in tests that are put on top of the original trend test.

The extremely large false-negative rates in the above and other subsequent simulated situations caused by the nature (low cancer rates and small group sample sizes) of a carcinogenicity experiment reinforce the important arguments that it is necessary to assume an overall false-positive rate of about 10% to raise the power (or to reduce the false-negative rate) of an individual statistical test that uses a very small multiplicity adjusted level of significance and that it is a big concern about the use of the joint test with the levels of significance recommended for the separate tests in the 2001 draft guidance document in the determination of the carcinogenicity of a new drug. Again, the producer's risk in a trend test alone is known (0.5% for a common tumor and 2.5% for a rare tumor in a two-species study) and is small in relation to the consumer's risk that can be 100 or 200 times of the level of the known producer's risk. The levels of significance recommended in the 2001 draft guidance for industry document were developed with the consideration of those situations in which the carcinogenicity experiment has great limitations. Trying to cut down the producer's risk (false-positive rate in regulatory review of toxicology studies) beyond that which safeguards against the huge consumer's risk (false-negative rates in regulatory review of toxicology studies) by the use of the joint test with the improper levels of significance is not consistent with the principles of the statistical science.

We also made a comparison of the false-negative rates (and levels of power) of the trend test alone between the NTP (Dinse 1985) and our simulation study based on the Weibull distribution as a quality control check of the results of our study. Table 8.4 contains the false-negative rates and levels of power (1—false-negative rate) under the various simulation conditions from the two studies. Although the same Weibull distribution models were used to generate the survival and tumor data, and 10,000 replicates (datasets) were used in both studies, there are some differences in experimental design as mentioned previously and in used levels of significance in statistical tests between the two studies. The levels of power of the NTP study are from the column under "Z_{NTP} and No continuity correction" in Table 3 (page 760) of Dinse (1985).

Again there are differences in design between NTP and our studies. There were only three treatment groups (control, low and high) and a total of 150 animals used in the NTP study, while there were four treatment groups (control, low, medium, and high) and a total of 200 animals used in our study. In our study, the 0.005 level of significance was used in the test if the background rate of the generated sample data of the control group is 1% or greater; otherwise, the 0.025 level of significance was used. The 0.05 level of significance was used in the NTP study.

Because of the differences in experimental design and level of significance used, an exact comparison of the results between the two studies cannot be made. However, the comparison of the results between the two studies has assured us that our simulation study results are very consistent with those from the NTP study. Our study has higher false-negative rates than those of the NTP study in all the simulation conditions except the following four (a) no dose effect on mortality, low background rate, tumors appear

Table 8.4 Comparison of false-negative rates (and levels of power) between NTP and our study based on the Weibull distribution

Simulation factors and conditions with factors				NTP study using 0.05 level of significance(1) (3)		OB/CDER/FDA study using either 0.005 or 0.025 level of significance(1) (2)	
Dose effect on mortality	Background tumor rate	Tumor appearance time	Dose effect on tumor prevalence	Power	False negative	Power	False negative
None	Low	Early	Small	0.46	0.54	0.37	0.63
None	Low	Late	Small	0.43	0.57	0.37	0.63
None	High	Early	Small	0.29	0.71	0.16	0.84
None	High	Late	Small	0.26	0.74	0.16	0.84
Small	Low	Early	Small	0.42	0.58	0.30	0.70
Small	Low	Late	Small	0.39	0.61	0.28	0.72
Small	High	Early	Small	0.25	0.75	0.12	0.88
Small	High	Late	Small	0.23	0.77	0.11	0.89
Large	Low	Early	Small	0.32	0.68	0.16	0.84
Large	Low	Late	Small	0.28	0.72	0.13	0.87
Large	High	Early	Small	0.19	0.81	0.06	0.94
Large	High	Late	Small	0.17	0.83	0.04	0.96
Average				0.31	0.69	0.19	0.81
None	Low	Early	Large	0.86	0.14	0.87	0.13
None	Low	Late	Large	0.83	0.17	0.86	0.14
None	High	Early	Large	0.65	0.35	0.66	0.34
None	High	Late	Large	0.62	0.38	0.62	0.38
Small	Low	Early	Large	0.81	0.19	0.78	0.22
Small	Low	Late	Large	0.76	0.24	0.73	0.27
Small	High	Early	Large	0.60	0.40	0.54	0.46
Small	High	Late	Large	0.55	0.45	0.46	0.54
Large	Low	Early	Large	0.64	0.36	0.46	0.54
Large	Low	Late	Large	0.56	0.44	0.36	0.64
Large	High	Early	Large	0.44	0.56	0.26	0.74
Large	High	Late	Large	0.38	0.62	0.16	0.84
Average				0.64	0.36	0.56	0.44

Notes (1) There are differences in design between NTP and OB/CDER/FDA studies. There were only three treatment groups (control, low, and high) and a total of 150 animals used in the NTP study, while there were four treatment groups (control, low, medium, and high) and a total of 200 animals used in the OB/CDER/FDA study. (2). In the OB/CDER/FDA study, the 0.005 level of significance was used in the test if the background rate of the generated sample data of the control group is 1% or greater; otherwise, the 0.025 level of significance was used. The 0.05 level of significance was used in the NTP study. (3). The levels of power of the NTP study are from the column under "Z_{NTP} and No continuity correction" in Table 3 (page 760) of Dinse (1985)

early, and large dose effect on tumor prevalence, (b) no dose effect on mortality, low background rate, tumors appear late, and large dose effect on tumor prevalence, (c) no dose effect on mortality, high background rate, tumors appear early, and large dose effect on tumor prevalence, and (d) no dose effect on mortality, high background rate, tumors appear late, and large dose effect on tumor prevalence. In those four simulation conditions, the false-negative rates from the NTP study are only a little bit higher than the corresponding rates from our study. The reason behind the above observation of the false-negative rates from the two studies is that the NTP study used the larger 0.05 level of significance in the trend test than that used in our study in which the level of significance 0.005 or 0.025 was used depending on the background rate of the control group of the generated tumor data. As mentioned previously, holding other factors that affect the power (or false-negative rate) constant, a larger level of significance used in a test will produce a lower false-negative rate.

Another observation of the false-negative rates from the NTP and our first studies is that the reductions in false-negative rate in the NTP study using a much larger level of significance (0.05) than that used in our study (0.005 or 0.025) are not as large as they might be expected to be if holding other factors constant. The reason behind this observation is that the total numbers of animal used (number of animals per group multiplied by a number of groups) in the two studies are different. The NTP study used the old study protocol that had only three treatment groups (control, low, and high) and 150 animals, while our study used the more recent study protocol that had four treatment groups (control, low, medium, and high) and 200 animals. The smaller total number of animals used in the NTP study cuts down the effect of reducing the false-negative rates caused by using a larger level of significance.

We did some additional runs of the simulation and took a closer look at the four exceptional cases mentioned above in which the false-negative rates from our first simulation study were equal or slightly lower than those of the NTP study. We found an explanation for those exceptional cases. All the four exceptional cases involved no dose effect on mortality, and large dose effect on the tumor prevalence. For all survival and tumor generated datasets, most of the calculated p values from the trend tests of 10,000 individual datasets are very small; that is, the trend tests will be considered as statistically significant regardless of the use of the level of significance of 0.05 used in the NTP study or of 0.005 or 0.025 used in our study.

8.3.2 Results of Our Second Simulation Study Based on the Gamma Distribution

The results of our second simulation study based on the Gamma distribution are presented in Table 8.5. As the estimated false-negative rates of Columns 5, 6, and 8 of this table show, our second simulation study yields consistent results with those of our first simulation study. Like in our first simulation study based on the Weibull distribution, the magnitude of the inflation of false-negative rate resulting from the

use of the joint tests requiring statistical significant results at the levels of significance recommended in the 2001 draft guidance document for the separate tests depends on all the four factors considered in the simulation study.

Results of our second simulation study based on the Gamma distribution show even larger inflations of the false-negative rate than those from the first simulation study based on the Weibull distribution. The inflations are most serious in the situations in which the dose has a large effect on tumor prevalence. The inflation can be as high as 204.5% (i.e., more than three times) of that when the trend test alone is required to be statistically significant to consider that the effect is statistically significant. In the first simulation study, most serious inflations also occur when the dose effect on death is small in addition to the large effect of dose on tumor prevalence. However, in the second simulation study, the dose effect on death extends from only small effect in the first simulation study to both none and small effects in this second study. The most serious inflation, 204.5%, occurs when the dose effect on tumor prevalence is large and the dose effect on death is none.

8.4 Recommended Decision Rules for Interpretations of Study Results of Various Types of Submissions and Various Statistical Tests

In rodent carcinogenicity studies, the extent needed for adjusting for the effect of multiple tests (controlling of the overall false-positive error) depends on statistical tests performed (trend tests alone, or control-high pairwise comparison tests alone, or joint tests of trend and control-high pairwise comparisons simultaneously), and studies conducted and included in an application submission (two two-year studies in two species, or a two-year study in only one species, or a combination of one two-year study in one species and a short- or medium-term study in the other species under ICH guideline).

Statistical procedures have been proposed for controlling the overall false-positive rate (Fairweather et al. 1998; Lin and Ali 2006; Lin 2010; Lin et al. 2010, 2016; Rahman and Lin 2009; Thomson and Lin 2009). In those publications, the statistical decision rules for controlling the overall false-positive rates associated with trend tests and pairwise comparisons separately and jointly for interpreting the final results of carcinogenicity studies in the three types of submissions are discussed. The recommended decision rules have been developed based on our empirical studies using historical control data of CD rats and CD mice (strains that are most widely used in studies of pharmaceuticals), and some strains (models) of transgenic mice, and on our simulation studies, including those discussed and included in the first part of this book chapter, to achieve an overall false-positive rate of around 10% for the various combinations of the statistical tests performed and the three different types of submissions.

Table 8.5 Estimated false-negative rates of trend test alone and trend test along with pairwise comparisons in simulations based on Gamma distribution

Dose effect on death	Background tumor rate	Tumor appear-ance	Dose effect on tumor rate	Trend	Trend and high	Trend and any	Percent change T-high	Percent change T-any
None	0.0527	Early	None	0.9880	0.9970	0.9940	0.9109	0.6073
None	0.0527	Early	Small	0.6480	0.7500	0.7390	15.74	14.04
None	0.0527	Early	Large	0.1540	0.4690	0.3380	204.5	119.5
None	0.0674	Late	None	0.9910	0.9980	0.9970	0.7064	0.6054
None	0.0674	Late	Small	0.7390	0.8180	0.8070	10.69	9.202
None	0.0674	Late	Large	0.2220	0.4940	0.3820	122.5	72.07
None	0.1937	Early	None	0.9970	0.9980	0.9980	0.1003	0.1003
None	0.1937	Early	Small	0.7770	0.8340	0.8180	7.336	5.277
None	0.1937	Early	Large	0.3400	0.5440	0.4650	60.00	36.76
None	0.1805	Late	None	0.9950	0.9960	0.9960	0.1005	0.1005
None	0.1805	Late	Small	0.8390	0.8830	0.8760	5.244	4.410
None	0.1805	Late	Large	0.2940	0.5220	0.4270	77.55	45.24
Small	0.0527	Early	None	0.9830	0.9980	0.9960	1.526	1.322
Small	0.0527	Early	Small	0.7180	0.8560	0.8460	19.22	17.83
Small	0.0527	Early	Large	0.1900	0.5030	0.3980	164.7	109.5
Small	0.0674	Late	None	0.9880	0.9990	0.9980	1.113	1.012
Small	0.0674	Late	Small	0.7370	0.8970	0.8860	21.71	20.22
Small	0.0674	Late	Large	0.2940	0.6410	0.5810	118.0	97.62
Small	0.1937	Early	None	0.9960	0.9970	0.9970	0.1004	0.1004
Small	0.1937	Early	Small	0.8370	0.9250	0.9100	10.51	8.722
Small	0.1937	Early	Large	0.3640	0.5910	0.5390	62.36	48.08
Small	0.1805	Late	None	0.9940	1.000	0.9990	0.6036	0.5030
Small	0.1805	Late	Small	0.8780	0.9780	0.9740	11.39	10.93
Small	0.1805	Late	Large	0.3840	0.7460	0.6980	94.27	81.77
Large	0.0527	Early	None	0.9900	1.000	1.000	1.010	1.010
Large	0.0527	Early	Small	0.8130	0.9390	0.9270	15.50	14.02
Large	0.0527	Early	Large	0.3760	0.6540	0.6220	73.94	65.43
Large	0.0674	Late	None	0.9910	1.000	1.000	0.9082	0.9082
Large	0.0674	Late	Small	0.8580	0.9850	0.9740	14.80	13.52
Large	0.0674	Late	Large	0.5810	0.9450	0.9220	62.65	58.69
Large	0.1937	Early	None	0.9980	1.000	1.000	0.2004	0.2004
Large	0.1937	Early	Small	0.8850	0.9910	0.9850	11.98	11.30

(continued)

Table 8.5 (continued)

Dose effect on death	Background tumor rate	Tumor appearance	Dose effect on tumor rate	Trend	Trend and high	Trend and any	Percent change T-high	Percent change T-any
Large	0.1937	Early	Large	0.5640	0.8710	0.8320	54.43	47.52
Large	0.1805	Late	None	0.9990	1.000	1.000	0.1001	0.1001
Large	0.1805	Late	Small	0.9270	1.000	1.000	7.875	7.875
Large	0.1805	Late	Large	0.6270	0.9970	0.9820	59.01	56.62

(Note of Table 8.5): The estimated false-negative rates under simulation numbers 1, 4, 7, 10, 13, 16, 19, 22, 25, 28, 31, and 34 are actually one minus the estimated false-positive rates because the assumption of no dose effect on tumor prevalence rate is used in those simulations

As mentioned previously, it is well known that, for a multi-group study (e.g., three doses and control), trend tests are more powerful than pairwise comparisons between the control and individual treated groups. Therefore, tests for a positive trend instead of pairwise comparison tests for a positive increase between control and high dose groups should be the primary tests in the evaluation of drug-related increases in tumor rate. However, also as mentioned previously, in order to further reduce the false-positive rate, some practicing pharmacologists/toxicologists have used the joint test with levels of significance recommended for the separate tests in the 2001 draft guidance document in interpreting a statistical significance of a test of drug effect in the development of an individual tumor types in a given tissue of tested animals. In the joint test, the drug effect on the development of a given tumor in a given tissue is considered as a statistical significance only if both the trend test and the C-H pairwise comparison of the tumor incidences are both simultaneously significant at the above-mentioned levels of significance for the separate tests for trend test alone and the separate C-H pairwise comparison test alone. We have recommended a new set of more appropriate levels of significance in terms of the balance between the false-positive rate and the false-negative rate for the joint tests for the three types of IND and NDA submissions.

In terms of studies conducted, most new drug application submissions consist of two two-year (chronic) carcinogenicity studies in two species. However, there are two situations in which a submission includes only a chronic carcinogenicity study conducted in one species.

In the first of the two above-mentioned situations, drug sponsors follow the International Conference on Harmonization (ICH) (1988) guideline S1B Testing for Carcinogenicity of Pharmaceuticals (1998) in the carcinogenicity evaluations of their new drugs. This ICH guideline outlines experimental approaches to the evaluation of carcinogenic potential that may obviate the need for the routine use of two long-term rodent carcinogenicity studies and allows for the alternative approach of conducting one long-term rodent carcinogenicity study together with a short- or medium-term rodent test. The short- or medium-term rodent test systems include such studies

as initiation-promotion in rodents, transgenic rodents, or newborn rodents, which provide rapid observations of carcinogenic endpoints in vivo. The special issue of Toxicologic Pathology (2001) published a large collection of papers on those short- or medium-term rodent test systems. In most submissions under the ICH guideline, drug sponsors conduct 26-week studies in transgenic mice as the short- or medium-term rodent test systems in the other species. Different strains (models) of transgenic mice have been used in the alternative carcinogenicity studies of pharmaceuticals. P53 ±, Tg.AC, TgrasH2, and XPA −/− are the models having been used by drug companies.

The FDA 2001 draft guidance for industry document on chronic rodent carcinogenicity studies of pharmaceuticals does not cover the discussions of designs, and methods of data analysis and interpretation of study results of transgenic mouse studies. However, except the part of interpretation of study results, the recommended methods of statistical analysis for the type of submissions including two chronic studies in rats and mice are still applicable to analyze the data from transgenic mouse studies excluding those studies using Tg.AC mice that have been rarely used nowadays. It becomes necessary only to develop decision rules for the interpretation of results for this type of submissions.

The second situation includes application submissions of post-marketing commitment studies of approved drugs or application submissions of studies of new drugs under which carcinogenicity studies of the drug products in one species are usually considered sufficient based on pharmacological and toxicological justifications. FDA does not have clear regulations on this issue of situations in which this type of submission is acceptable. Drug sponsors are asked to discuss the issue and get agreements on the issue with the Agency in advance. As in the type of submissions under ICH guideline, the recommended methods of statistical analysis for the type of submissions with two chronic studies in rats and mice are still applicable to analyze the study data of this third type of submissions. However, it also just becomes necessary only to develop decision rules for the interpretation of results for this type of submissions.

8.4.1 Decision Rules for the Separate and Joint Tests in Submissions with Two Two-Year Studies in Two Species

In the past, CDER/FDA statistical reviewers of carcinogenicity studies used the statistical decision rule described in Haseman (1983, 1984a) in their tests for significance of positive trends in tumor incidence of a given tumor type. The decision rule was originally developed for pairwise comparison tests in tumor incidence between the control and the high dose groups and was derived from results of carcinogenicity studies of environmental compounds conducted at National Toxicology Program (NTP). Strains of Fischer 344 rats and B6C3F1 mice were used in the NTP studies.

Like most studies of pharmaceuticals, four treatment/sex groups with 50 animals in each group were used in those NTP studies. All of those NTP studies lasted for 2 years. The decision rule tests the significant differences in tumor incidence of a given tumor type between the control and the high dose groups at 0.05 significance level for rare tumors and at 0.01 significance level for common tumors. Again, a tumor type with a background rate of 1% or less is classified as rare by Haseman; more frequent tumors are classified as common. Haseman's original study and a second study using more recent data with higher background tumor rates show that the use of this decision rule in the control-high pairwise comparison tests would result in an overall false-positive rate between 7 and 8% based on earlier studies, and between 10 and 11% based on more recent studies, respectively (Haseman 1983, 1984a, b).

Concerns were raised that applying the levels of significance described by Haseman (1983) to analyses of positive trend tests would lead to an excessive overall false-positive error rate since data from all treatment groups are used in the tests and considerably lower background tumor rates in carcinogenicity studies of pharmaceutical compounds than those NTP studies of environmental compounds can yield a wrongly significant result. Results from studies conducted within and outside CDER/FDA show that these concerns were valid. Based on studies conducted by us and by NTP, the overall false-positive error resulting from interpreting trend tests by the use of the above decision rules is about twice as large as that associated with control-high pairwise comparison tests (Lin and Rahman 1998).

Based on studies using real historical control data of CD mice and CD rats from Charles River Laboratories and on simulation studies conducted internally in CDER/FDA and in collaboration with NTP, new statistical decision rules for tests for a positive trend alone in tumor incidence of a tumor type in this type of submissions have been developed. These new decision rules test the positive trends in incidence rates in rare and common tumors at 0.025 and 0.005 levels of significance, respectively. The new decision rules achieve an overall false-positive rate of around 10% in a standard submission with two two-year studies in two species (Lin 1995; Lin and Rahman 1998; Rahman and Lin 2008). The 10% overall false-positive rate is considered by CDER/FDA as appropriate in a regulatory setting of reviewing carcinogenicity studies of new drugs.

As mentioned above, statistical literature emphasizes methods for testing for positive trends alone in tumor incidence rate. There are, as mentioned previously, situations, however, in which pairwise comparisons alone between control and individual treated groups may be more appropriate than the trend tests. Under those situations, the decision rules described in Haseman (1983) should be used in interpreting the results of the control-high pairwise comparison tests. The Haseman decision rules recommend that the control-high pairwise comparison alone be tested at 0.01 and 0.05 significance levels common and rare tumors, respectively.

As also mentioned above, the joint test often used by pharmacologists/toxicologists in the determination of a statistical significance of a drug effect on the development of an individual tumor type in a given tissue requires both the trend test and the C-H pairwise comparison of the incidences of the tumor type to

be both simultaneously significant at the levels of significance for the separate tests recommended in the 2001 draft guidance document.

Because the nature of the carcinogenicity studies included in this type of submissions, i.e., with the small group sizes (50–70 animals/group) used in regular chronic carcinogenicity studies as a surrogate of a big population of mice or rats, and with low tumor incidence rate endpoints, the false-negative rate is already inherently big. Therefore, it is necessary to assume larger overall false-positive rates, such as 0.1 (10%), in a carcinogenicity study than those used in other types of drug development studies, such as clinical trials, to reduce the large false-negative rate (or to increase the low power of detecting a true effect) inherent in the nature of the studies mentioned above.

Based on the important results of our new simulation studies discussed in the first part this chapter, and the results of our previously published empirical studies, the following recommended decision rules for submissions with two two-year studies in two species have been developed for the joint test using both the trend test and the C-H pairwise comparison simultaneously in the interpretation of a statistically significant result of a drug effect on the development of an individual tumor type. It is recommended that the significance levels 0.005 and 0.025 in the trend test and 0.05 and 0.10 in the C-H pairwise comparison for common and rare tumors, respectively, be used in the joint test in each of the two simultaneous tests in this type of submissions. The use of the newly recommended decision rules will still result in about 10% overall false-positive rate in a submission with two chronic studies in two species.

The newly developed set of decision rules are to be used as alternative rules for the joint test in the determination of a statistically significant effect of a drug on the development of a given tumor type. The development of this newly recommended alternative set of decision rules for the joint test should not be wrongly construed as the invalidation of the previously developed and recommended decision rules for the trend tests alone and for the C-H pairwise comparisons alone.

The new decision rules for the joint test are designed to produce about the same levels of false-positive rate in a joint test as those produced by a trend test alone as shown in the sampled plot in Fig. 8.3. Therefore, the use of the newly recommended decision rules in the joint tests in a standard submission with two chronic studies in mice and rats will also result in an overall false-positive rate about 10% based on the results of Lin and Rahman (1998) and Rahman and Lin (2008).

People may have the opinion that the 0.1 level of significance we have recommended for rare tumors for the C-H pairwise comparison component of the joint test is too excessive. We have taken a close look into this issue by evaluating the false-positive rates in the small range between 0.0 and 0.01 of background tumor rates where our recommended level of significance of 0.1 for rare tumors in the C-H pairwise comparison component of the joint test produced much smaller false-positive rates than those from the trend test alone. Figure 8.4 shows the comparison of attained false-positive rates of the joint test using the newly recommended levels of significance with those of the trend test alone using the levels of significance rec-

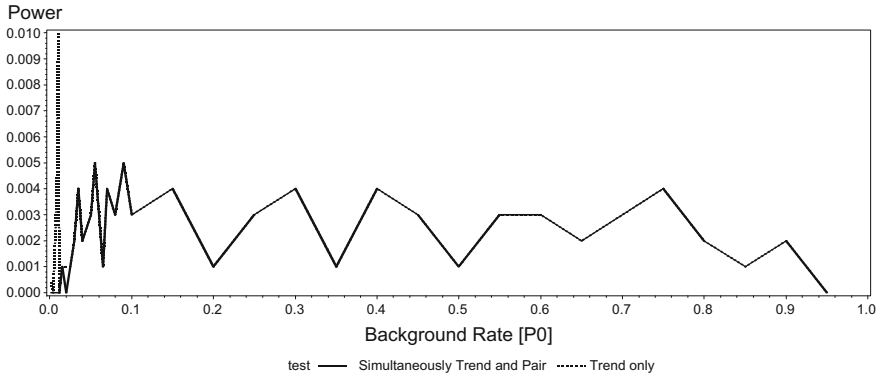


Fig. 8.3 Plot of attained false-positive rates of the trend test alone and the joint test using test levels of 0.1 for rare tumors and 0.05 for common tumors for the C-H pairwise comparison component

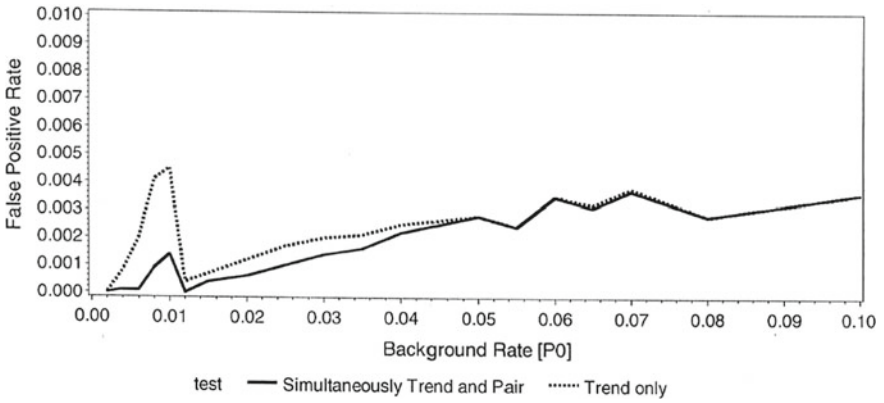


Fig. 8.4 Plot of attained false-positive rates of the trend test alone and the joint test using test levels of 0.1 for rare tumors and 0.05 for common tumors

ommended in the 2001 guidance for industry document for background tumor rates between 0.0 and 0.1.

After looking at Fig. 8.4, we feel that the recommended 0.1 level of significance for rare tumors for the C-H pairwise component of the joint test is still not large enough to reduce the large differences in false-positive rate between the joint test and the trend test alone within the (0.0, 0.01) range of background tumor rates. We have tried larger levels of significance of 0.15, 0.2, and 0.0.25 for rare tumors for the C-H component of the joint test. We have obtained the identical plots of attained false-positive rates of the joint test using those larger levels of significance for the C-H comparison component and of the trend test alone in the background rates between 0.0 and 0.01 as shown in Fig. 8.5. The reasons for the identical plots for the three levels of significance 0.15, 0.2, and 0.25 for rare tumors for the C-H component of the joint test are that the estimated variances of the distributions of the test statistics

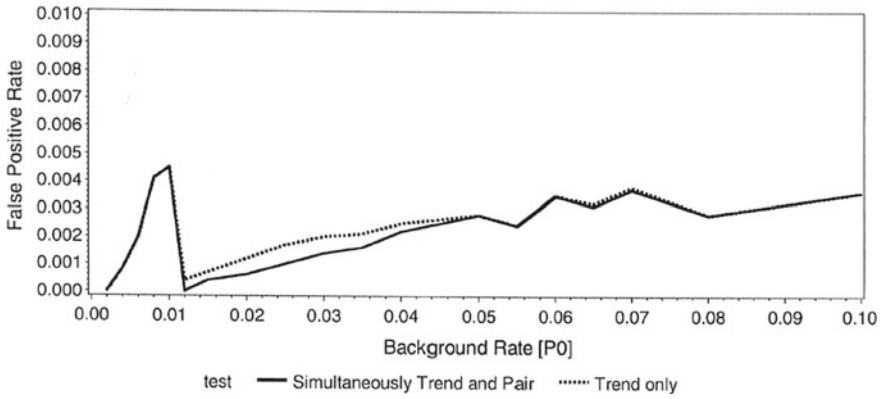


Fig. 8.5 Plot of attained false-positive rates of the trend test alone and the joint test using test levels of 0.15 or 0.2 or 0.25 for rare tumors and 0.05 for common tumors

become so small with those very low background tumor rates and that the p values of the C-H pairwise component and of the test for positive trend based on those distributions are smaller than 0.15 and 0.025, respectively.

The reason for the small p values of the Peto prevalence asymptotic trend test discussed above can be easily seen by looking at the normal approximation test statistic $Z = T/[V(T)]^{1/2}$ where $T = \sum_i D_i\ (O_i - E_i)$, $V(T) = \sum_i \sum_j D_i\ D_j\ V_{ij}$, D_i is the dose level of the ith group, $O_i = \sum_k O_{ik}$, $E_i = \sum_k E_{ik}$, $V_{ij} = \sum_k V_{ijk}$, $V_{ijk} = \alpha_k P_{ik}(\delta_{ij} - P_{jk})$, $\alpha_k = O_{.k}(R_k - O_{.k})/(R_k - 1)$, $\delta_{ij} = 1$ if $i = j$, and $= 0$, otherwise, $E_{ik} = O_{.k}\ P_{ik}$, $O_{.k} = \sum_i O_{ik}$, R_k is the number of animals that have not died of the tumor type of interest, but come to autopsy in time interval k, P_{ik} is the proportion of R_k in group I, O_{ik} is the observed number of autopsied animals in group i and interval k found to have the incidental tumor type, and $O_{.k} = \sum_i O_{ik}$. For definitions of the above mathematical notations and more detailed discussions on the test method, readers are referred to the FDA 2001 draft guidance for industry document (U.S. Department of Health and Human Services 2001). When the background rates under null hypothesis of no drug effect are low (less than 0.01), all the estimated variances, $V_{ijk,}$ $V_{ij,}$ and V(T) will be small, and the calculated Z statistics will be large, and the p-values will be small.

It is noted that the levels of attained false-positive rates of the test for positive response alone in Figs. 8.4 and 8.5 are not at the same levels as those in Fig. 8.3 in the range of 0.0–0.01 of background rate. The attained false-positive rates of the test for positive response alone in the range of background rates are higher in Fig. 8.3 than those in Figs. 8.4 and 8.5 for the background rates between 0.0 and 0.01. Figures 8.4 and 8.5 used data generated from a different simulation study using an exact instead of the asymptotic prevalence method in both the separate test for positive response alone and the joint test for the positive response and for the C-H pairwise increase comparison in the computations of attained false-positive rates. The exact method

for the Peto prevalence trend test described in Mehta et al. (1988) was used along with the Fisher exact test in this different simulation study.

The higher attained false-positive rates between 0.0 and 0.01 background tumor rates in Fig. 8.3 than those in Figs. 8.4 and 8.5 are related to the issue of underestimation of p values of the test for positive response based on asymptotic normal approximation used in calculations of the attained false-positive rates of Fig. 8.3. The generated tumor-bearing animals of the four treatment groups based on those small background rates of only 50 animals each will be all small under the null hypothesis of no carcinogenic effect of the drug. It is well known (Ali 1990) that p values of the Peto asymptotic test for positive response will be underestimated when the total tumor-bearing animals across all treatment groups are small. Under those situations, it is recommended to use exact permutation tests (Gart et al. 1986; U.S. Department of Health and Human Services 2001; Rahman et al. 2016; Lin et al. 2016) to replace the asymptotic tests to correct the problem. But as mentioned in Sect. 8.2, the age-adjusted prevalence asymptotic method described in Peto et al. (1980) was used in our simulations discussed in that section. Exact permutation tests will result in larger p values than those from the asymptotic tests, and, therefore, yield lower false-positive rates as demonstrated in Figs. 8.4 and 8.5 based on the generated data from the different simulation and on the use of an exact method of the Peto prevalence test. The conservative exact permutation tests will, in turn, result in increasing false-negative rates. We feel that the recommended 0.1 level of significance for rare tumors for the C-H pairwise comparison component of the joint test is justified even when exact permutation tests for positive response are used in the comparisons of attained false-positive rates of the joint test and the test for positive response alone in our simulation studies.

We have also evaluated the performance, in terms of the control of false-negative rates, of the newly recommended levels of significance for the joint test based on datasets generated from the Weibull distributions described in Sect. 8.2.1. The estimated false-negative rates from the evaluations are presented in Table 8.6. This table contains also estimated false-negative rates (Columns 7 and 8, 10 and 11) when two other pairs of levels of significance (0.05, 0.05) and (0.05, 0.01) in addition to the newly recommended pair of levels of (0.1, 0.05) were used for the control-high pairwise comparisons (Columns 6 and 9) in the joint tests.

Rates of Columns 5 and 8 of Table 8.6 are slightly different from those of Columns 6 and 7 of Table 8.3 because only 2,000 replications were used this evaluation while 10,000 replications were used in the evaluation described in Sects. 8.2.1 and 8.3.1.

The rates of Column 5 (basing on the trend test alone) and Columns 6 and 9 (basing on the joint test using the newly recommended levels of significance) and Columns 8 and 11 (basing on the joint test with the levels of significance for the separate tests recommended in the 2001 draft guidance document) show that the use of the joint test with the newly recommended levels of significance greatly reduces the false-negative rates to about the levels of using the trend test alone in all but 1 of the 36 simulation conditions conducted. In the simulation condition (#18), the reductions in inflation in false-positive rate were still significantly reduced by two thirds (from 144.6% (or 153.3% in Table 8.3) to 58.70%).

Table 8.6 Estimated false-negative rates from trend test alone and the application of the new decision rules in the joint test basing data generated from Weibull distributions

Dose effect on death	Tumor appearance	Dose effect on tumor rate	Background tumor rate	Trend 025_0 05	Trend and high 10_05	Trend and high 05_05	Trend and high 05_01	Percent change Tr-high 10_05	Percent change Tr-high 05_05	Percent change Tr-high05_01
No	Early	No	0.0500	0.9840	0.9875	0.9925	0.9925	0.3557	0.8638	0.8638
No	Early	Small	0.0500	0.6200	0.6220	0.6220	0.6995	0.3226	0.3226	12.82
No	Early	Large	0.0500	0.1290	0.1315	0.1315	0.1745	1.938	1.938	35.27
No	Late	No	0.0501	0.9850	0.9870	0.9920	0.9925	0.2030	0.7107	0.7614
No	Late	Small	0.0501	0.6250	0.6285	0.6290	0.7035	0.5600	0.6400	12.56
No	Late	Large	0.0501	0.1380	0.1405	0.1405	0.1915	1.812	1.812	38.77
No	Early	No	0.2000	0.9940	0.9945	0.9945	0.9975	0.0503	0.0503	0.3521
No	Early	Small	0.2000	0.8235	0.8275	0.8275	0.8720	0.4857	0.4857	5.889
No	Early	Large	0.2000	0.3235	0.3270	0.3270	0.3960	1.082	1.082	22.41
No	Late	No	0.2001	0.9955	0.9955	0.9955	0.9985	0.0000	0.0000	0.3014
No	Late	Small	0.2001	0.8310	0.8385	0.8385	0.8885	0.9025	0.9025	6.919
No	Late	Large	0.2001	0.3625	0.3700	0.3700	0.4645	2.069	2.069	28.14
Small	Early	No	0.0500	0.9810	0.9920	0.9960	0.9960	1.121	1.529	1.529
Small	Early	Small	0.0500	0.7015	0.7430	0.7470	0.8500	5.916	6.486	21.17
Small	Early	Large	0.0500	0.2075	0.2370	0.2370	0.4005	14.22	14.22	93.01
Small	Late	No	0.0501	0.9795	0.9980	0.9995	0.9995	1.889	2.042	2.042
Small	Late	Small	0.0501	0.7390	0.8385	0.8520	0.9210	13.46	15.29	24.63
Small	Late	Large	0.0501	0.2760	0.4380	0.4400	0.6750	58.70	59.42	144.6

(continued)

Table 8.6 (continued)

Dose effect on death	Tumor appearance	Dose effect on tumor rate	Background tumor rate	Trend 025_0 05	Trend and high 10_05	Trend and high 05_05	Trend and high 05_01	Percent change Tr-high 10_05	Percent change Tr-high 05_05	Percent change Tr-high05_01
Small	Early	No	0.2000	0.9935	0.9970	0.9970	1.000	0.3523	0.3523	0.6543
Small	Early	Small	0.2000	0.8635	0.9095	0.9095	0.9685	5.327	5.327	12.16
Small	Early	Large	0.2000	0.4740	0.5675	0.5675	0.7445	19.73	19.73	57.07
Small	Late	No	0.2001	0.9955	0.9990	0.9990	1.000	0.3516	0.3516	0.4520
Small	Late	Small	0.2001	0.8805	0.9705	0.9705	0.9955	10.22	10.22	13.06
Small	Late	Large	0.2001	0.5180	0.8180	0.8180	0.9400	57.92	57.92	81.47
Large	Early	No	0.0500	0.9865	0.9990	1.000	1.000	1.267	1.368	1.368
Large	Early	Small	0.0500	0.8465	0.9220	0.9320	0.9585	8.919	10.10	13.23
Large	Early	Large	0.0500	0.5215	0.6335	0.6370	0.8050	21.48	22.15	54.36
Large	Late	No	0.0501	0.9825	1.000	1.000	1.000	1.781	1.781	1.781
Large	Late	Small	0.0501	0.8650	0.9860	0.9955	0.9960	13.99	15.09	15.14
Large	Late	Large	0.0501	0.6540	0.9515	0.9615	0.9735	45.49	47.02	48.85
Large	Early	No	0.2000	0.9965	0.9995	0.9995	1.000	0.3011	0.3011	0.3512
Large	Early	Small	0.2000	0.9325	0.9935	0.9935	0.9995	6.542	6.542	7.185
Large	Early	Large	0.2000	0.7475	0.9395	0.9395	0.9830	25.69	25.69	31.51
Large	Late	No	0.2001	0.9955	1.000	1.000	1.000	0.4520	0.4520	0.4520
Large	Late	Small	0.2001	0.9585	1.000	1.000	1.000	4.330	4.330	4.330
Large	Late	Large	0.2001	0.8325	0.9995	0.9995	1.000	20.06	20.06	20.12

The results in Table 8.6 show that, in this type of submissions, the use of the newly recommended levels of significance for the joint test in the data generated from the Weibull distribution has greatly reduced the false-negative rates when compared to those presented in Table 8.3 (and also in Table 8.5 based on data generated from the Gamma distribution) resulting from the use of the levels of significance recommended for the separate tests in the 2001 draft guidance in the joint test. As mentioned above, the false-negative rates from the using the newly recommended levels of significance in the joint test have been reduced to about the levels of those using the levels of significance recommended in the 2001 guidance for industry document in the trend test alone. Due to the space limitation, the newly recommended levels of significance for the joint test were not evaluated on the data generated from the Gamma distribution described in Sect. 8.2.2. However, we expect to see the similar results of great reductions in the inflation of false-negative rates as in those the above evaluation using data generated from the Weibull distribution.

8.4.2 Decision Rules for the Separate and Joint Tests in Submissions Under ICH Guideline with a Combination of One Two-Year Study in Rats and a Transgenic Mouse Study

False positives therefore arise primarily from the two-year rodent carcinogenicity study in submissions under the ICH guideline because of more common tumors among a large number of tumor/organ combinations tested than those in the transgenic mouse study or other short- or medium-term studies. In the transgenic mouse study or other short- or medium-term studies that use much small number of animals, the false-positive rate will not be inflated even when multiple statistical tests on different tumors are performed because tumor background rates are very low, and the number of tumor types developed in the tested animals is small. In studies using Tg.AC transgenic mice, only data of incidence rates and weekly numbers of skin papillomas are used to test the carcinogenic effect of the drug. In those studies, a positive result is almost surely a true instead of a false-positive. This means that if each of the individual tests is performed at a given level of significance (e.g. 0.05), the overall false-positive rate will be still close to the given level. Therefore, there is no or little need to adjust the effect of multiple tests.

Based on a large number of empirical and simulation studies conducted within CDER/FDA mentioned in the previous section, the following decision rules are recommended for controlling the overall false-positive rate of submissions under the ICH guideline.

For tests for positive trend alone, it is recommended that common and rare tumors are tested at 0.005 and 0.025 significance levels, respectively, in the two-year study and at 0.05 and 0.05 significance levels, respectively, in the alternative study.

For control-high pairwise comparison alone, it is recommended that common and rare tumors are tested at 0.01 and 0.05 significance levels, respectively, in the two-year study and at 0.05 and 0.05 significance levels, respectively, in the alternative study.

For tests for positive trend and control-high pairwise comparison jointly, it is recommended that common and rare tumors are tested at 0.005 and 0.025 significance levels, respectively, in trend test, and at 0.05 and 0.10 significance levels, respectively, in control-high pairwise comparison in the two-year study and at 0.05 and 0.05 significance levels, respectively, in both trend test and control-high pairwise comparison in the alternative study for both common and rare tumors.

The use of the above decision rules in a submission under the ICH guideline will result in an overall false-positive rate about 5% in the two-year study and another overall false-positive rate also around 5% for the short- or medium-term studies in the tests for trend and control-high group pairwise comparisons, separately or jointly. This will result in an overall false-positive rate of around 10% for the entire submission.

8.4.3 Decision Rules for the Separate and Joint Tests in Submissions with Only One Two-Year Study in One Species

For tests for trend and control-high group comparisons, separately or jointly, the overall false-positive rate is a function of the number of tests performed, background tumor rate, group sizes, and variability of the population data. Holding other factors unchanged, the overall false rate increases, in general, as the number of tests performed increases. The number of tests performed in a submission with only one two-year study in one species is about half of that of a submission with two two-year studies in two species. Higher levels of significance should be used in the multiplicity adjustment in the submission with only one two-year study in one species.

Also based on a large number of empirical and simulation studies conducted within CDER/FDA mentioned in the previous section, the following decision rules are recommended for the separate and joint tests in controlling the overall false-positive rate in submissions with only one two-year study in one species justified under the special regulatory conditions mentioned previously.

For tests for positive trend alone, it is recommended that common and rare tumors are tested at 0.01 and 0.05 significance levels, respectively.

For control-high pairwise comparison alone, it is recommended that common and rare tumors are tested at 0.025 and 0.10 significance levels, respectively.

For tests for positive trend and control-high pairwise comparison jointly (i.e., the joint test), it is recommended that common and rare tumors are tested at 0.01 and 0.05 significance levels, respectively, in trend test, and at 0.05 and 0.10 significance levels, respectively, in control-high pairwise comparison.

Table 8.7 Recommended decision rules (levels of significance) for controlling the overall false-positive rates for various statistical tests performed and for various types of submissions

	Tests for positive trend alone	Control-high pairwise comparison alone	Tests for positive trend and control-high pairwise comparison jointly
Standard two-year studies with two species and two sexes	Common and rare tumors are tested at 0.005 and 0.025 significance levels, respectively	Common and rare tumors are tested at 0.01 and 0.05 significance levels, respectively	Common and rare tumors are tested at 0.005 and 0.025 significance levels, respectively, in trend test, and at 0.05 and 0.10 significance levels, respectively, in control-high pairwise comparison
Alternative ICH studies (one two-year study in one species and one short- or medium-term alternative study, two sexes)	Common and rare tumors are tested at 0.005 and 0.025 significance levels, respectively, in the two-year study, and at 0.05 and 0.05 significance levels, respectively, in the alternative study	Common and rare tumors are tested at 0.01 and 0.05 significance levels, respectively, in the two-year study, and at 0.05 and 0.05 significance levels, respectively, in the alternative study	Common and rare tumors are tested at 0.005 and 0.025 significance levels, respectively, in trend test, and at 0.05 and 0.10 significance levels, respectively, in control-high pairwise comparison in the two-year study, and at 0.05 and 0.05 significance levels, respectively, in both trend test and control-high pairwise comparison in the alternative study
Standard two-year studies with one species only and two sexes	Common and rare tumors are tested at 0.01 and 0.05 significance levels, respectively	Common and rare tumors are tested at 0.025 and 0.10 significance levels, respectively	Common and rare tumors are tested at 0.01 and 0.05 significance levels, respectively, in trend test, and at 0.05 and 0.10 significance levels, respectively, in control-high pairwise comparison

The use of the above decision rules in a submission with only one two-year study will result in an overall false-positive rate about 10% in the tests for trend and control-high group pairwise comparisons, separately or jointly.

The recommended decision rules (levels of significance) for controlling the overall false-positive rates for various statistical tests performed and types of submissions discussed above are summarized in Table 8.7.

Acknowledgements The authors would like to express their thanks to Dr. Yi Tsong, Director of Division of Biometrics 6, Office of Biostatistics, CDER/FDA, for encouraging them to publish the results of their regulatory research studies.

References

Ali, M. W. (1990). Exact versus asymptotic tests of trend of tumor prevalence in tumorigenicity experiments: A comparison of P-values for small frequency of tumors. *Drug Information Journal, 24,* 727–737.

Dinse, G. E. (1985). Testing for a trend in tumor prevalence rates: I. nonlethal tumors. *Biometrics, 41,* 751–770.

Fairweather, W. R., Bhattacharyya, A., Ceuppens, P. P., Heimann, G., Hothorn, L. A., Kodell, R. L., et al. (1998). Biostatistical methodology in carcinogenicity studies. *Drug Information Journal, 32,* 401–421.

Gart, J. J., Krewski, D., Lee, P. N., Tarone, R. E., & Wahrendorf, J. (1986). *Statistical methods in cancer research, Volume III—the design and analysis of long-term animal experiments*. International Agency for Research on Cancer, World Health Organization.

Haseman, J. K. (1983). A reexamination of false-positive rates for carcinogenesis studies. *Fundamental and Applied Toxicology, 3,* 334–339.

Haseman, J. K. (1984a). Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies. *Environmental Health Perspective, 58,* 385–392.

Haseman, J. K. (1984b). Use of historical control data in carcinogenicity studies in rodents. *Toxicologic Pathology, 12,* 126–135.

International Conference on Harmonization (ICH). (1988). Guideline S1B, *Testing for Carcinogenicity of Pharmaceuticals*.

Lin, K. K. (1988). Peto prevalence method versus regression methods in analyzing incidental tumor data from animal carcinogenicity experiments: An empirical study. In *1988 American Statistical Association Annual Meeting Proceedings (Biopharmaceutical Section)*, New Orleans, LA.

Lin, K. K. (1995). A regulatory perspective on statistical methods for analyzing new drug carcinogenicity study data. *Bio/Pharam Quarterly, 1*(2), 18–20.

Lin, K. K. (2000a). Carcinogenicity studies of pharmaceuticals. In S. C. Chow (Ed.), *Encyclopedia of biopharmaceutical statistics* (pp. 88–103). New York: Marcel Dekker.

Lin, K. K. (2000b). A progress report on the guidance for industry for statistical aspects of the design, analysis, and interpretation of chronic rodent carcinogenicity studies of pharmaceuticals. *Journal of Biopharmaceutical Statistics, 10,* 481–501.

Lin, K.K. (2010). Carcinogenicity studies of pharmaceuticals. In S. C. Chow (Ed.), *Encyclopedia of Biopharmaceutical Statistics* (3rd ed.), Revised and Expanded, Vol. 1. New York and London: Informa Healthcare.

Lin, K. K., & Ali, M. W. (2006). Statistical review and evaluation of animal carcinogenicity studies of pharmaceuticals. In C. R. Buncher & J. Y. Tsay (Eds.), *Statistics in the pharmaceutical industry* (3rd ed.). New York: Chapman & Hall/CRC.

Lin, K. K., Jackson, M. T., Min, M., Rahman, M. A., & Thomson, S. F. (2016). Recent research projects by the FDA's pharmacology and toxicology statistics team. In Lanju Zhang (Ed.), *Nonclinical biostatistics for pharmaceutical and biotechnology industries*. New York: Springer.

Lin, K. K., & Rahman, M. A. (1998). Overall false positive rates in tests for linear trend in tumor incidence in animal carcinogenicity studies of new drugs. *Journal of Pharmaceutical Statistics, with discussions, 8*(1), 1–22.

Lin, K. K., Thomson, S. F., & Rahman, M. A. (2010). The design and statistical analysis of toxicology studies. In G. Jagadeesh, S. Murthy, Y. Gupta, & A. Prakash (Eds.), *Biomedical research: From ideation to publications* (1st ed.). India: Wolters Kluwer.

Mehta, C. R., Patel, N. R., & Wei, L. J. (1988). Constructing exact significance tests with restricted randomization rules. *Biometrika, 75*(2), 295–302.

Peto, R., Pike, M. C., Day, N. E., Gray, R. G., Lee, P. N., Parish, S., et al. (1980). Guidelines for simple, sensitive significance tests for carcinogenic effects in long–term animal experiments. In *IARC monographs on Long-term and Short-term Screening Assays for Carcinogens: An critical appraisal*. Geneva: World Health Organization.

Rahman, A. M., & Lin, K. K. (2008). A comparison of false positive rates of Peto and poly-3 methods for long-term carcinogenicity data analysis using multiple comparison adjustment method suggested by Lin and Rahman. *Journal of Biopharmaceutical Statistics, 18*(5), 849–858.

Rahman, A. M., & Lin, K. K. (2009). Design and analysis of chronic carcinogenicity studies of pharmaceuticals in rodents. In K. E Peace, (Ed.), *Design and analysis of clinical trials with time-to-event endpoints*. Boca Raton, FL: Chapman & Hall/CRC, Taylor & Francis Group, LLC.

Rahman, M. A., & Lin, K. K. (2010). Statistics in pharmacology. In G. Jagadeesh, S. Murthy, Y. Gupta, & A. Prakash (Eds.), *Biomedical research: From ideation to publications* (1st ed.). India: Wolters Kluwer (India).

Rahman, M. A., Lin, K. K., Tiwari, R. C., & Jackson, M. (2016). Exact Poly-*K* Test. *Communications in Statistics—Simulation and Computation, 45*(7), 2257–2266.

Society of Toxicologic Pathology. (2001). *Toxicologic pathology* (Vol. 29) (Supplement).

Thomson, S. F., & Lin, K. K. (2009). Design and summarization, analysis and interpretation of time-to-tumor response in animal studies: A Bayesian Perspective. In K. E. Peace, (Ed.), *Design and analysis of clinical trials with time-to-event endpoints*. Boca Raton, FL, London, and New York: Chapman & Hall/CRC, Taylor & Francis Group, LLC.

U.S. Department of Health and Human Services. (2001). *Guidance for industry, statistical aspects of the design, analysis, and interpretation of chronic rodent carcinogenicity studies of pharmaceuticals*. Maryland, U.S.: Center for Drug Evaluation and Research, Food and Drug Administration, U.S. Department of Health and Human Services.

Chapter 9
A Prematurely Halted, Randomized, Controlled Clinical Trial of Alendronate Treatment in Patients with Gaucher Disease

Shumei S. Sun

9.1 Introduction

Gaucher disease is an autosomal recessive disease caused by severely decreased intracellular hydrolysis of glucosylceramides and other glucosphingolipids. Nearly all cases are due to a heritable deficiency of lysosomal glucocerebrosidase (GC) that causes massive accumulation of unhydrolyzed glucosylceramides and other glucosphingolipids within macrophages of the liver, spleen, bone marrow, and other tissues (Beutler and Grabowski 2001). Most symptomatic patients with GC deficiency are treated with periodic injections of recombinant GC (Barton et al. 1991). This therapy results in gratifying decreases in glucosylceramide storage in the liver and spleen, but signs and symptoms relating to bone disease, including bone pain and osteopenia, have either required years of therapy for improvement, or have been completely refractory to enzyme therapy (Elstein et al. 1996; Rosenthal et al. 1995).

The purpose of this investigation was to determine whether the osteopenia that is seen in most adults with Gaucher disease can be corrected by bone anti-resorptive adjunctive therapy in patients who are receiving enzyme therapy (Harinck et al. 1984; Ciana et al. 1997; Ostiere et al. 1991). We initiated a two-year, double-blind, two-arm, placebo-controlled trial of alendronate, at a dose of 40 mg/day, in adults with Gaucher disease who had been treated for at least six months with GC enzyme therapy. Therapeutic outcome was monitored by dual-energy x-ray absorptiometry (DXA) measurements of bone mineral density (BMD) and bone mineral content (BMC) at the lumbar spine and whole body DXA scans at entry into the study and at six-month intervals until the end of the 24-month study.

S. S. Sun (✉)
Virginia Commonwealth University, Richmond, VA, USA
e-mail: shumei.sun@vcuhealth.org; ssun@vcu.edu

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_9

Table 9.1 Number of examinations by number of patients

Number of examinations	Number of patients		
	Alendronate	Placebo	Total
Lumbar spine scan	16	15	31
1	6	7	13
2	5	5	10
3	2	1	3
4	0	1	1
5	3	1	4
Whole body scan	15	15	30
1	5	7	12
2	5	5	10
3	2	1	3
4	1	1	2
5	2	1	3

9.2 Summary Statistics

Data were collected on 31 patients. The study was halted prematurely according to widely accepted stopping rules (O'Brien and Fleming 1979) because of the development of significant positive differences between group assignments. When the group randomization code was broken, it was found that 16 patients had been randomized to alendronate and 15 to placebo. Only four patients completed all of the study examinations with measurements at baseline, 6, 12, 18, and 24 months. Among the 27 patients who did not finish the study, 13 had one DXA measurement at the lumbar spine and 12 had one whole body DXA scan. Ten of the patients had two DXA measurements at the lumbar spine and two had whole body DXA scans. It was found that one-half of these ten patients had been assigned to alendronate and one-half to placebo. Table 9.1 presents the number of patients by number of measurements in the alendronate and placebo groups, separately and together.

9.3 Randomization at Baseline

To test the randomization of patients at entry into the trial, baseline means and standard deviations of sex, age, lumbar spine BMC and BMD, total body BMC and BMD, total body fat mass, and total body lean mass were calculated in each of the two groups. A two-sample t test was used to determine whether there were significant differences in these variables between the groups at baseline. None of the differences noted between the randomly assigned groups were significant at entry into the study as shown in Table 9.2.

Table 9.2 Baseline mean bone mineral density, body fat mass, body lean mass, and percent fat mass in alendronate group and placebo group

	Alendronate		Placebo		P value
	N	Mean ± Std	N	Mean ± Std	
Lumbar spine scan					
Age (years)	16	39.62 ± 7.60	15	34.85 ± 8.99	0.1203
BMC_TOT (g)	16	58.44 ± 15.85	15	61.37 ± 13.43	0.5849
BMD_TOT (g/cm^2)	16	0.946 ± 0.145	15	0.955 ± 0.140	0.8539
Whole body scan					
Age (years)	15	39.27 ± 7.73	15	34.85 ± 8.99	0.1598
BMC_TOT (g)	15	2379.0 ± 391.43	15	2516.5 ± 549.99	0.4366
BMD_TOT (g/cm^2)	15	1.137 ± 0.133	15	1.171 ± 0.148	0.5165
Total body fat mass (g)	15	18032 ± 3672	15	19251 ± 9723	0.6551
Total body lean mass (g)	15	49717 ± 12933	15	49407 ± 10519	0.9431
(%) Fat	15	0.2609 ± 0.0525	15	0.2590 ± 0.0860	0.9419

9.3.1 Statistical Analyses

For BMC and BMD of the lumbar spine, total body BMC (BMC_TOT) and total body BMD (BMD_TOT), total body fat mass, and total body lean mass, the differences between baseline values at entry into the study and measurements six months later were calculated for each patient. We used paired t tests to determine whether the differences in measurements of these parameters differed from zero, and we used two-sample t tests to determine whether measurements of these parameters at six months after entry into the study, when compared to baseline values, showed significant differences in the alendronate and placebo groups. Table 9.3 presents these results.

Ten patients in the alendronate group and eight patients in the placebo group had more than one measurement (including baseline measurements), and five patients in alendronate group and three in the placebo group had more than two measurements (including baseline). Due to the small sample sizes, a longitudinal model was not applied to the dataset. Instead, we calculated the mean of the measured variables between the second visit and the most recent visit and the differences between this mean and baseline measurements for each individual. A paired t test was used to check whether the resulting differences were qualitatively different from zero, and a two-sample t test was applied to determine whether bone density improved in patients assigned to alendronate compared to those assigned to placebo. The results of this analysis are shown in Table 9.4.

In a third statistical approach, linear regressions were run to obtain the change in slopes of the outcome parameters for patients who had more than one visit (Wenstrup et al. 2004). The slopes indicate the pattern of change over time in the measured

Table 9.3 Differences between baseline and six-month examination values for lumbar and whole body DXA scans in alendronate and placebo groups

	Alendronate (P-value)	Placebo (P-value)	2-Sample
	N = 10	N = 8	P value
Lumbar spine scan			
BMC_TOT (g)	2.5181 ± 2.4415^{++}(0.0098)	−0.4580 ± 2.2013(0.5747)	0.0164*
BMD_TOT (g/cm^2)	0.0313 ± 0.0240^{++}(0.0026))	0.0012 ± 0.0282(0.9106)	0.0263*
Whole body scan			
BMC_TOT (g)	12.85 ± 45.54(0.3956)	15.58 ± 59.17(0.4805)	0.9129
BMD_TOT (g/cm^2)	0.0034 ± 0.0138(0.4510)	0.0109 ± 0.0208(0.1819)	0.3738
Total body fat mass (g)	−0.759 ± 1997(0.9991)	307.51 ± 1832(0.6494)	0.7402
Total body lean mass (g)	−16.82 ± 742.0(0.9444)	668.91 ± 913.5(0.0771)	0.0975
(%) Fat	−0.002 ± 0.0227(0.8093)	0.0022 ± 0.0192(0.7534)	0.6965

$^{+}$0.01 < P value < 0.05, significantly different from 0 at 0.05 level
$^{++}$P value < 0.01, significantly different from 0 at 0.01 level
*0.01 < P value < 0.05, significantly different between having alendronate and placebo at 0.05 level
**P value < 0.01, significantly different between having alendronate and placebo at 0.01 level

Table 9.4 Differences between baseline and the mean of the rest of the examinations (mean of rest—baseline) for alendronate and placebo groups

	Alendronate (P-value)	Placebo (P-value)	2-Sample
	N = 10	N = 8	P value
Lumbar spine scan			
BMC_TOT (g)	3.3018 ± 2.3514^{++}(0.0016)	−0.7130 ± 1.7579(0.2892)	0.0010**
BMD_TOT (g/cm^2)	0.0380 ± 0.0244^{++}(0.0008)	0.0026 ± 0.0291(0.8067)	0.0126*
Whole body scan			
BMC_TOT (g)	9.49 ± 50.93(0.5701)	13.86 ± 54.58(0.4958)	0.8630
BMD_TOT (g/cm^2)	0.0070 ± 0.0125(0.1093)	0.0114 ± 0.0185(0.1255)	0.5616
Total body fat mass (g)	145.9 ± 2259(0.8427)	274.7 ± 1832(0.6843)	0.8980
Total body lean mass (g)	−118.8 ± 358.9(0.7482)	586.1 ± 867.5(0.0976)	0.1670
(%) Fat	0.0008 ± 0.0245(0.9215)	0.0023 ± 0.0197(0.7521)	0.7775

$^{+}$0.01 < P value < 0.05, significantly different from 0 at 0.05 level
$^{++}$P value < 0.01, significantly different from 0 at 0.01 level
*0.01 < P value < 0.05, significantly different between having alendronate and placebo at 0.05 level
**P value < 0.01, significantly different between having alendronate and placebo at 0.01 level

Table 9.5 Differences in mean slopes of bone mineral content and bone mineral density, total body fat mass, total body lean mass and percent body fat between alendronate and placebo groups per month

	Alendronate (P-value)	Placebo (P-value)	2-Sample
	N = 10	N = 8	P value
Lumbar spine scan			
BMC_TOT (g/mo)	3.641 ± 2.874^{++}(0.0031)	−1.596 ± 3.124(0.1918)	0.0020**
BMD_TOT (g/cm^2/mo)	0.0415 ± 0.0301^{++}(0.0018)	−0.005 ± 0.0379(0.6977)	0.0097**
Whole body scan			
BMC_TOT (g/mo)	11.94 ± 63.40(0.0597)	6.36 ± 77.15(0.1941)	0.8682
BMD_TOT (g/cm^2/mo)	0.0132 ± 0.0194(0.5663)	0.0173 ± 0.0341(0.8224)	0.7496
Total body fat mass (g/mo) mo)6mo(g/6mo momo mo mo) (g/6mo) mo (g/6mo (g/mo) mo)	−491.3 ± 3531(0.6703)	−67.8 ± 2329(0.9367)	0.7745
Total body lean mass (g/mo) (g/mo (g/6mo mo)	−170.2 ± 1035(0.6155)	964.1 ± 1830(0.1799)	0.1158
(%) Fat (%/mo)	−0.008 ± 0.040(0.5259)	−0.001 ± 0.026(0.8784)	0.6806

$^{+}$0.01 < P value < 0.05, significantly different from 0 at 0.05 level
$^{++}$P value < 0.01, significantly different from 0 at 0.01 level
*0.01 < P value < 0.05, significantly different between having alendronate and placebo at 0.05 level
**P value <0.01, significantly different between having alendronate and placebo at 0.01 level.

parameters. Table 9.5 shows the two-sample t test results of the mean slopes for the alendronate and placebo groups.

9.4 Results

As shown in Table 9.3, lumbar spine BMC and BMD increased significantly in the alendronate group after the first six months of treatment compared to baseline. The mean increases and standard deviations were 2.5181 ± 2.4415 grams and 0.0313 ± 0.0240 grams per centimeter squared, for BMC_TOT and BMD_TOT, respectively. In the placebo group, however, BMC_TOT decreased and BMD_TOT increased slightly after the first six months on placebo compared to baseline. The mean changes and standard deviations were −0.4580 ± 2.2013 grams and 0.0012 ± 0.0282 grams per centimeter squared for BMC_TOT and BMD_TOT, respectively. Compared to the placebo group, the alendronate group had significantly greater increases in BMC_TOT and BMD_TOT after six months of treatment. In terms of the whole body DXA scan, BMC_TOT, BMD_TOT, total body fat mass, and total

body lean mass did not show any significant differences between the alendronate and placebo groups. However, patients assigned to placebo had a greater total body fat mass and greater total body lean mass compared to the patients assigned to alendronate, but the percent body fat did not differ significantly between the two groups.

Table 9.4 shows results similar to those in Table 9.3 but with a more salient increase in mean lumbar spine BMC_TOT and BMD_TOT after six months of treatment with alendronate compared to baseline. The mean change in BMC_TOT was 3.3018 ± 2.3514 grams after six months on alendronate versus −0.72130 ± 1.7579 grams after six months on placebo. Changes of BMD_TOT over six months were 0.0380 ± 0.0244 grams per centimeter squared in the alendronate group and 0.0026 ± 0.0291 grams per centimeter squared in the placebo group. The increases in mean BMC_TOT and BMD_TOT from baseline during treatment were significantly greater for patients assigned to alendronate than for patients assigned to placebo.

Table 9.5 shows the pattern of change over time. Mean lumbar spine BMC_TOT and BMD_TOT increased significantly over time in the alendronate group but decreased in the placebo group. The average rates of change in BMC_TOT were 3.641 ± 2.874 grams per six months in the group assigned to alendronate and −1.596 ± 3.124 grams per six months in the group assigned to placebo. The mean rates of change in BMD_TOT were 0.0415 ± 0.0301 grams per centimeter squared per six months in patients assigned to alendronate and –0.005 ± 0.0379 grams per centimeter squared in patients assigned to placebo.

9.5 Conclusion

The mean bone mineral content and bone mineral density of the lumbar spine increased significantly in patients assigned to alendronate rather than placebo over a period of time as little as six months. However, the mean BMC_TOT and BMD_TOT by whole body DXA scan did not show significant differences between the alendronate and placebo groups. Due to the small sample sizes in the dataset, the random effects longitudinal model could not be applied to the data to detect differences between the two groups. Of potential clinical importance, the dataset at the cessation of the study suggests that patients assigned to placebo but not to alendronate gained in percent body fat and lean body mass.

References

Barton, N. W., Brady, R. O., Dambrosia, J. M., et al. (1991). Replacement therapy for inherited enzyme deficiency: macrophage-targeted glucocerebrosidase for Gaucher's disease. *The New England Journal of Medicine*, *324*, 1464–1470.

Beutler, E., & Grabowski, G. (2001). Gaucher disease. In C. R. Scriver, A. L. Beaudet, D. Valle, et al. (Eds.), *The metabolic and molecular basis of inherited disease* (8th ed., pp. 3635–3638) New York, NY: McGraw-Hill.

Ciana, G., Cuttini, M., & Bembi, B. (1997). Short-term effects of pamidronate in patients with Gaucher's disease and severe skeletal involvement [letter]. *The New England Journal of Medicine, 337,* 712.

Elstein, D., Hadas-Halpern, I., Itzchaki, M., Lahad, A., Abrahamov, A., Zimran, A. (1996). Effect of low-dose enzyme replacement therapy on bones in Gaucher disease patients with severe skeletal involvement. *Blood Cells, Molecules and Diseases*, *22,* 101–108.

Harinck, H. I., Bijvoet, O. L., van der Meer, J. W., Jones, B., & Onvlee, G. J. (1984). Regression of bone lesions in Gaucher's disease during treatment with aminohydroxypropylidene bisphosphonate [letter]. *Lancet, 2,* 513.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 35,* 49–56.

Ostiere, L., Warner, T., Meunier, P. J., et al. (1991). Treatment of type 1 Gauchet's disease affecting bone with aminohydroxypropylidene bisphosphonate (pamidronate). *The Quarterly Journal of Medicine*, *79*, 503–515.

Rosenthal, D. I., Doppell, S. H., Mankin, H. J., et al. (1995). Enzyme replacement therapy for Gaucher disease: Skeletal responses to macrophage-targeted glucocerebrosidase. *Pediatrics*, *96*, 629–627.

Wenstrup, R. J., Bailey, L., Grabowski, G. A., Moskovitz, J., Oestreich, A. E., Wu, W., et al. (2004). Gaucher disease: Alendronate sodium improves bone mineral density in adults receiving enzyme therapy. *Blood, 104*,1253–1257.

Chapter 10
Mediation Modeling in Randomized Trials with Non-normal Outcome Variables

Jing Cheng and Stuart A. Gansky

10.1 Introduction

In many health studies, the intervention is designed to change some post-randomization (intermediate) variable, such as knowledge, attitudes, behavior, biomarkers or social factors, so that the change in the intermediate variable will lead to improvement in the final health outcomes of interest (MacKinnon and Luecen 2011). Such an intermediate variable is usually called mediator, explaining how and/or why an exposure/program/treatment changes an outcome of interest. For example, the Detroit Dental Health Project's Motivational Interviewing DVD (DDHP MI-DVD) trial was a randomized dental trial of a Motivational Interviewing (MI) intervention to prevent early childhood caries (ECC) in low income African-American children (0–5 years) in Detroit, Michigan (Ismail et al. 2011). In the study, caregivers in both intervention and control groups watched a 15-min education video on children's oral health. The control group (DVD only) was then provided general recommendations on diet, oral hygiene, and dental visits. For the intervention group (MI+DVD), a MI interviewer reviewed the child's dental examination with caregivers and discussed caregivers' personal thoughts and concerns about specific goals for their child's oral health. A brochure with caregivers' specific goals was then printed and placed in a convenient place at home. Families in the MI+DVD group also received booster calls within 6 months of the intervention. The study hypothesized that the MI+DVD intervention would change the caregivers' and children's behaviors in oral hygiene and then the behavioral changes would lead to improved oral health in children. In these studies, researchers are not only interested if the intervention works but also if and how much the intervention affects the outcome through and around the intermediate mediator.

J. Cheng · S. A. Gansky (✉)
University of California, San Francisco, USA
e-mail: Stuart.Gansky@ucsf.edu

J. Cheng
e-mail: Jing.Cheng@ucsf.edu

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_10

The effect of an intervention through the mediator is called the indirect or mediation effect, showing that the intervention affects the outcome through the intermediate variables as designed. The effect around the mediator is called the direct effect, indicating that the intervention changes the outcome directly or involving some other intermediate variables in a heretofore undiscovered mechanism. Knowing those effects helps us to better understand the working mechanism of an intervention such that future programs can tailor specific program components to target specific important mediators and consequently lead to bigger improvement in health outcomes.

Many conventional and causal mediation approaches (Baron and Kenny 1986; Cole and Maxwell 2003; Daniels et al. 2012; Goetgeluk et al. 2008; Imai et al. 2010a; Jo et al. 2011; MacKinnon 2008; Pearl 2001; Robins and Greenland 1992; Rubin 2004; Small 2012; Sobel 1982, 2008; Steyer et al. 2014; Ten Have et al. 2007; Van der Laan and Petersen 2008; VanderWeele and Vansteelandt 2009) have been developed for continuous outcomes. When the outcome of interest is a non-continuous but binary, multinomial, or count outcome, mediation approaches relying on linear models may not be appropriate. For non-continuous outcomes such as binary outcomes with nonlinear models, MacKinnon and Dwyer (1993) showed that the traditional product and difference methods give different results. Therefore, we will discuss mediation approaches for non-continuous outcomes in this chapter.

10.2 Traditional Mediation Analysis

Since Wright (1920) developed a path analysis as a special case of structural equation modeling (SEM), various SEM-based methods have been developed for mediation analysis (e.g., Judd and Kenny 1981; Sobel 1982; MacKinnon 2008). Baron and Kenny (1986) provided an excellent conceptual description of mediation more than 30 years ago. Since then, their proposed approach on mediation in the context of linear models has been widely used in many areas for mediation analyses on continuous outcomes. Figure 10.1 illustrates Baron and Kenny's approach using the SEM approach and involving multi-step regressions, where Z_i denotes the randomized

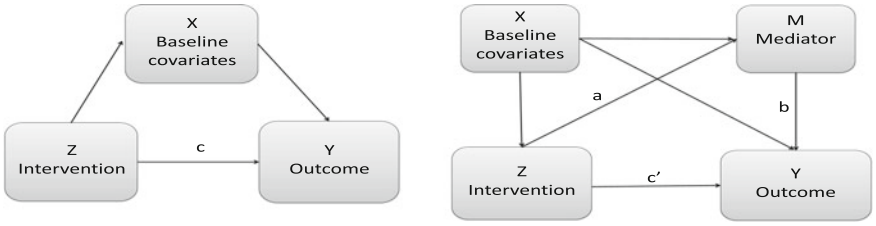


Fig. 10.1 Baron and Kenny's mediation approach

treatment, M_i for the observed mediator level, X_i for the observed baseline covariates, and Y_i for observed outcome for subject i.

- Total effect of treatment on outcome (c) is estimated by regressing the outcome variable on the treatment: $Y_i = \gamma_1 + \kappa_1 X_i + cZ_i$
- Regress the mediator on the treatment controlling for baseline covariates: $M_i = \gamma_2 + \kappa_2 X_i + aZ_i$
- Direct effect of treatment on outcome (c′) is estimated by regressing the outcome variable on the treatment while controlling for the mediation: $Y_i = \gamma_3 + \kappa_3 X_i + bM_i + c'Z_i$
- Check if all the coefficients a, b, and c are significant and if c' is smaller than c. If the criteria are met, then the mediation effect of the treatment on the outcome via the mediator is estimated by $(c - c')$.

That is, Baron and Kenny's approach estimates the mediation effect as the difference between total effect (c) and direct effect (c'). Alternatively the mediation effect can be estimated as a product of the two coefficients a and b (MacKinnon 2008). MacKinnon et al. (1995) showed that the mediation effect by the product method is equivalent to the effect by the difference method for continuous outcomes with linear models.

Although not explicitly expressed, the traditional mediation approaches require a series of ignorability assumptions to have a causal interpretation of direct and mediation effects, such as no unmeasured confounding of the Z-M, Z-Y, and M-Y relationships and no confounders of the M-Y relationship that are affected by the treatment Z (Pearl 2001; VanderWeele and Vansteelandt 2009). We will discuss the ignorability assumptions further under the potential outcome framework and introduce recently developed approaches relaxing some of those assumptions in the next sections. Although Baron and Kenny's approach may have low statistical power in some situations (MacKinnon et al. 2002) and their criteria may not be met sometimes (Holmbeck 2002), the estimated direct and mediation effects have causal interpretations under assumptions that the linear models and a series of ignorability assumptions hold, and there is no treatment and mediator interaction. Several causal methods (Pearl 2001; Jo 2008; Sobel 2008; VanderWeele and Vansteelandt 2009; Imai et al. 2010b) have described the SEM approach under the counterfactual potential outcome framework and relaxed the assumption of no treatment by-mediator interaction. For example, VanderWeele and Vansteelandt (2009) provided closed-form formula for the direct and mediation effects allowing treatment by-mediator interaction in linear models.

When the mediator and outcome are not continuous such that linear models do not fit the data, generalized linear models have been used for the mediation analysis with a logit or probit link for binary data or log link for count data:

$$\begin{aligned} g(E[Y_i|Z_i]) &= \gamma_1 + cZ_i \\ h(E[M_i|Z_i]) &= \gamma_2 + aM_i \\ g(E[Y_i|Z_i, M_i]) &= \gamma_3 + bM_i + c'Z_i. \end{aligned}$$

However, MacKinnon and Dwyer (1993) illustrated that the estimated mediation effect based on the product (ab) and the difference ($c - c'$) can be very different in those nonlinear models due to different scales that are used. After standardization, the estimated mediation effect based on the product and difference would be much closer (MacKinnon and Dwyer 1993; Coxe and MacKinnon 2010). However, the standardized product or difference methods may still lead to biased estimate for the causal mediation effect in some situations (Imai et al. 2010a; Pearl 2012). Additionally the results from those traditional approaches depend on specific statistical models because they do not provide a general definition of causal direct and mediation effects independent of specific statistical models before the analyses.

10.3 The Counterfactual Framework, Notation and Assumptions

Different from traditional approaches, recently developed causal mediation approaches adopt the potential outcome framework (Neyman 1923; Rubin 1974) and define causal direct and indirect (mediation) effects independent of a specific statistical model (Valeri and VanderWeele 2013). In this chapter, we will first conceptually define causal direct and indirect (mediation) effects without reference to a specific statistical model and then discuss different statistical models to identify and estimate those causal effects under appropriate assumptions. We will first consider the total, direct, and indirect (mediation) effects on the additive scale as a difference of two potential outcomes, and we will then discuss effects on other scales (e.g., corresponding to logit for binary outcomes and log for count outcomes) in next sections.

We make the Stable Unit Treatment Value Assumption (SUTVA), which says that a subject's potential outcome is not related to the randomization or mediation value of other subjects or the method of administration of randomization or the mediator. The SUTVA assumption allows us use scalar indices rather than vector indices in potential variable notation and enables linking the potential variables to observed variables. Under SUTVA, we let M_i^z denote the potential value of a mediator under treatment $Z_i = z$ for subject i. In a two-arm study, M_i^z has two versions: M_i^1 under treatment and M_i^0 under control. However, in practice we are not able to observe both potential mediator values but only one of M_i^1 and M_i^0 depending on which treatment group subject i was actually assigned to. We use $Y_i^{z,m}$ to denote the potential outcome subject i would have under the treatment $Z_i = z$ and mediator $M_i = m$, and Y_i^{z,M_i^z} for potential outcome under $Z_i = z$. Below we will use $Y_i^{z,m}$ to define controlled effects and Y_i^{z,M_i^z} for natural effects. Again, only one version of multiple potential outcomes will be observed for a subject depending on the actual treatment and mediator value subject i had.

Under the potential outcome framework, we will define both the individual level causal effects and population average causal effects. However, in real studies, we are not able to observe all the potential mediators and potential outcomes under treatment

and control for a subject who would only take either treatment or control, so the individual level causal effects cannot be identified. On the other hand, the population average causal effects can be identified under some assumptions, such as different sets of assumptions from Pearl (2001), Robins (2003), Van der Laan and Petersen (2008), Hafeman (2009), and Imai et al. (2010a) on sequential ignorability of treatment and mediator, and the assumption of no treatment by-mediator interaction in some methods (Robins 2003; Hafeman 2009). Van der Laan and Petersen (2008), Imai et al. (2010a) and Ten Have and Joffe (2010) provide good reviews of assumptions to achieve nonparametric identifiability of the causal effects.

The total effect (TE) of the treatment for subject i and its average across subjects are, respectively,

$$TE_i = Y_i^{1,M_i^1} - Y_i^{0,M_i^0}, \quad \bar{TE} = E(Y_i^{1,M_i^1} - Y_i^{0,M_i^0}),$$

which is the total effect of the treatment ($Z = 1$) on outcome Y compared to control ($Z = 0$) no matter whether the effect is through or around the mediator M. Note that in some situations, we are interested in the effect conditional on baseline covariates $\bar{TE} = E(Y_i^{1,M_i^1} - Y_i^{0,M_i^0}|X_i)$. The total effect of the treatment has two components: the treatment effect around the mediator, called the direct effect, and the treatment effect through the mediator, called the indirect or mediation effect. There are two sets of definitions on these effects proposed in the literature (Pearl 2001; Ten Have et al. 2007; Imai et al. 2010a; Robins 2003; Ten Have and Joffe 2010): controlled and natural effects.

The controlled direct effect (CDE) of the treatment for subject i and its average across subjects while fixing the mediator at m are

$$CDE_i^m = Y_i^{1m} - Y_i^{0m}, \quad \bar{CDE}^m = E(Y_i^{1m} - Y_i^{0m}),$$

which is the treatment effect compared to control while fixing the mediator at m, and the controlled mediation effect (CME) of m versus m' when fixing the treatment z and its average are

$$CME_i^z = Y_i^{zm} - Y_i^{zm'}, \quad \bar{CME}^z = E(Y_i^{zm} - Y_i^{zm'}),$$

$$\text{for } z = 0, 1 \text{ and all } m \neq m',$$

which is the effect of mediator (at m vs. at m') on the outcome under treatment z. Conditional on baseline covariates X_i, the controlled direct and mediation effects are $\bar{CDE}^m = E(Y_i^{1m} - Y_i^{0m}|X_i)$ and $\bar{CME}^z = E(Y_i^{zm} - Y_i^{zm'}|X_i)$.

In contrast to the controlled effects for setting the mediator at a fixed level m, the natural effects set the mediator at its "natural" level that would be achieved under treatment z. The natural direct effect (NDE) of the treatment for subject i and its average across subjects when the mediator is set at its level under treatment z are

$$NDE_i^z = Y_i^{1,M_i^z} - Y_i^{0,M_i^z}, \quad N\bar{D}E_i^z = E(Y_i^{1,M_i^z} - Y_i^{0,M_i^z}),$$

which is the treatment effect on outcome compared to control while having the mediator at its potential level M_i^z, and the natural mediation (indirect) effect (NME) and its average when fixing treatment z are

$$NME_i^z = Y_i^{z,M_i^1} - Y_i^{z,M_i^0}, \quad N\bar{M}E^z = E(Y_i^{z,M_i^1} - Y_i^{z,M_i^0}),$$

which is the outcome change under treatment z that would be observed if the mediator would change from the value under control M_i^0 to the value under treatment M_i^1. Conditional on baseline covariates X_i, the natural direct and mediation effects are $N\bar{D}E_i^z = E(Y_i^{1,M_i^z} - Y_i^{0,M_i^z}|X_i)$ and $N\bar{M}E^z = E(Y_i^{z,M_i^1} - Y_i^{z,M_i^0}|X_i)$. In real studies, we may not be able to set the mediator at a specific level and therefore natural effects are probably preferred.

However, because the natural effects involve the counterfactual potential outcome $Y_i^{z,M_i^{z^*}}$ ($z \neq z^*$) corresponding to both levels (z and z^*) of Z, the identification of natural effects requires stronger assumptions than controlled effects. To identify the effects, different sets of assumptions from Pearl (2001), Robins (2003), Van der Laan and Petersen (2008), VanderWeele and Vansteelandt (2009), and Imai et al. (2010a) among others have been proposed, including different versions of sequential ignorability assumption of treatment and mediator. For example, Imai et al. (2010a, b) and Cheng et al. (2017) assumed

$$\{Y_i^{z^*,m}, M_i^z\} \perp Z_i|X_i = x; \quad Y_i^{z^*,m} \perp M_i^z|Z_i = z, X_i = x, \quad \text{for all } z, z^*, m. \tag{10.1}$$

This assumption says that (1) the treatment is independent of potential mediators and potential outcomes given the baseline covariates; and (2) the mediators are independent of the potential outcomes given the treatment and baseline covariates. In the MI+DVD study, the first ignorability assumption is reasonable because of the randomization of the treatment. However, the second ignorability assumption may not hold even in the randomized MI+DVD study because the mediator, oral health behavior, after randomization was not randomly assigned. Even though it is not guaranteed by randomization, the second ignorability assumption may hold after conditioning on baseline covariates and treatment. That is, the mediator, oral health behavior, was as if randomized among subjects in the same treatment group who have the same baseline characteristics.

Under sequential ignorability, the distribution of the counterfactual potential outcome is nonparametrically identified (Imai et al. 2010b; Pearl 2012):

$$f(Y_i^{z,M_i^{z^*}}|X_i = x) = \int_M f(Y_i|M_i = m, Z_i = z, X_i = x)dF_{M_i}(m|Z_i = z^*, X_i = x), \quad x \in X; z, z^* = 0, 1.$$

$$E[Y_i^{z,M_i^{z^*}}|X_i = x] = \sum_m E(Y_i|Z_i = z, M_i = m, X_i = x)P(M_i = m|Z_i = z^*, X_i = x) \tag{10.2}$$

That is, the distribution (expectation) of the counterfactual potential outcome on the left-hand side can be expressed as a function of the distribution of observed data on the right-hand side. This result of nonparametric identifiability of the average causal effects enables the estimation of the causal effects based on potential outcomes and mediators we do not observe.

10.4 Mediation Analysis on Binary Outcomes

Many conventional and causal mediation approaches have been developed for continuous outcomes. For non-continuous outcomes such as binary outcomes with nonlinear models, MacKinnon and Dwyer (1993) showed that the traditional product method and difference method generally give different results, even though the two methods are approximately equivalent when the binary outcome is rare under assumptions shown by VanderWeele and Vansteelandt (2010). By the mediation formula (10.2), the conventional product method provides a consistent estimate for causal direct and mediation effects when the linear models and ignorability assumptions hold for the data (Imai et al. 2010a; VanderWeele and Vansteelandt 2009). Imai et al. (2010b) assume a mediator model and an outcome model:

$$M_i^{Z_i} \sim f_M(\theta_M = h^{-1}(\alpha_M + \beta_M Z_i + \eta_M^T X_i)) \tag{10.3}$$

$$Y_i^{Z_i, M_i^{Z_i}} \sim f_Y(\theta_Y = g^{-1}(\alpha_Y + \beta_Y Z_i + \gamma_Y M_i^{Z_i} + \xi_Y Z_i M_i^{Z_i} + \eta_Y^T X_i)) \tag{10.4}$$

where the link functions h and g are monotonic and differentiable functions; e.g., logit or probit link for binary M_i or Y_i. Then they used Monte Carlo integration to compute the direct and indirect (mediation) effects from (10.2) for general outcomes, including binary outcomes when the linear models do not hold. Specifically, their approach first builds a mediator model and an outcome model based on observed data, then samples $M_i^{z^*}$ from the mediator model and samples the counterfactual potential outcome $Y_i^{z, M_i^{z^*}}$ from the outcome model, and next uses (10.2) to compute the direct and indirect (mediation) effects. The *mediation* package in R implements their approach for common types of mediators and outcomes, including binary outcomes.

While Imai et al.'s approach estimates the causal direct and mediation effects as risk differences between treatment and control for general outcomes including binary outcomes marginalized over the baseline covariates, VanderWeele and Vansteelandt (2010) considered causal direct and mediation effects on non-addictive scale such as direct effect and mediation effect odds ratios (ORs) specifically for binary outcomes. They considered a logit model (10.5), a special case of the natural effects models (Lange et al. 2012; Vansteelandt et al. 2012), for a binary outcome and then derived the natural direct effect odds ratio (10.6) and natural mediation effect odds ratio (10.7):

$$logit[E\{Y_i^{z,M_i^{z^*}}|X_i\}] = \theta_0 + \theta_1 z + \theta_2 z^* + \theta_3 X_i \tag{10.5}$$

$$\frac{\text{odds}\{Y_i^{z,M_i^{z^*}}|X_i\}}{\text{odds}\{Y_i^{z^*,M_i^{z^*}}|X_i\}} = exp[\theta_1(z - z^*)] \tag{10.6}$$

$$\frac{\text{odds}\{Y_i^{z^*,M_i^{z}}|X_i\}}{\text{odds}\{Y_i^{z^*,M_i^{z^*}}|X_i\}} = exp[\theta_2(z - z^*)] \tag{10.7}$$

The natural direct and mediation effect ORs are considered a more natural scale for binary outcomes than the additive scale in Imai et al.'s approach (VanderWeele and Vansteelandt 2010; Vansteelandt 2010; Loeys et al. 2013). The approach by Vansteelandt et al. (2012) does not require a mediator model but still requires an imputation outcome model built on observed data in treatment, mediator, and baseline covariates. Then the counterfactual outcome $Y_i^{z,M_i^{z^*}}$ is predicted based on the imputation outcome model when z and z^* are not equal. The approach next regresses observed and imputed counterfactual potential outcomes on z, z^*, and X by Model (10.5) and obtains the "simple imputation estimate" by (10.6) and (10.7). Note that as opposed to the marginal effect in Imai et al.'s approach, the imputation estimate based on the natural effect models provides estimated conditional effects and allows for evaluating moderation effects of covariates X_i. Both approaches by Imai et al. (2010a, b) and VanderWeele and Vansteelandt (2010) allow treatment by-mediator interaction.

Alternatively, Elliott et al. (2010) constructed principal strata (Frangakis and Rubin 2002; Rubin 2004; VanderWeele 2008; Gallop et al. 2009) based on the joint distribution of the potential mediator value under treatment and control and then developed a Bayesian approach to estimate the principal strata-specific intent-to-treat (ITT) effect of the treatment where the treatment has no effect on the mediator (called "disassociative effect") and does change the mediator ("associative effect"). The mediation effect is then a function of associative effects and probabilities associated with the binary mediator and binary outcome.

10.5 Mediation Analysis on Count and Zero-Inflated Count Outcomes

In addition to binary outcomes, the outcome variable in many studies is often a count following a Poisson or Negative Binomial distribution, or a zero-inflated count that has a higher probability of being zero than expected under a Poisson or Negative Binomial distribution, such as number of doctor or emergency visits, number of admissions and readmissions to a hospital, number of complications, and number of decayed, missing and filled primary teeth (dmft) or tooth surfaces (dmfs). In the DDHP MI+DVD study, the outcomes of interest are the number of new untreated cavities, dmft and dmfs at the end of the study 2 years later compared between MI+DVD and DVD-only groups (Ismail et al. 2011), which contain 24–62% zeros

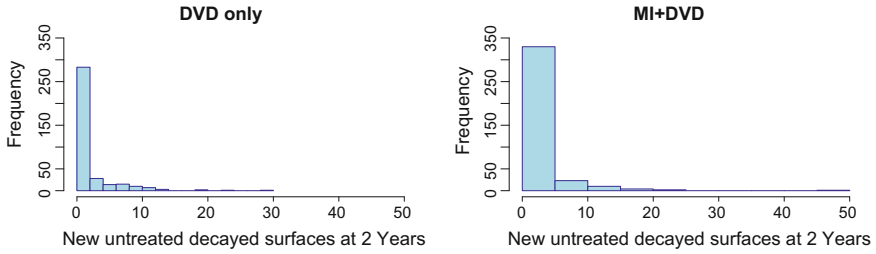


Fig. 10.2 Histograms of the numbers of new untreated decayed surfaces (cavities)

in various dental outcomes (Fig. 10.2) because the majority of the children did not have any new untreated cavities, dmft and dmfs at the end of the study.

10.5.1 Methods Under the Assumption of Sequential Ignorability

10.5.1.1 Mediation Analysis Without Post-baseline Confounders

As discussed above, we are not able to observe the counterfactual potential outcome involved in the natural effects $Y_i^{z,M_i^{z^*}}$ $(z \neq z^*)$ in a study. However, under sequential ignorability, the distribution of the counterfactual potential outcome $Y_i^{z,M_i^{z^*}}$ $(z \neq z^*)$ is nonparametrically identified (Imai et al. 2010a, b; Pearl 2012).

For a count outcome following a Poisson or Negative Binomial distribution, Imai et al.'s approach (2010b) fits a log linear outcome model (10.4) for the observed count outcome on observed treatment, mediator, and covariates, then samples the counterfactual outcome from the log linear outcome model, and then uses Monte Carlo integration to compute the direct and indirect (mediation) effects from the mediation formula (10.2) on the additive scale. Alternatively, Valeri and VanderWeele (2013) considered a log linear natural effect model for the count outcome and provided formula for the direct and indirect (mediation) effects on the rate ratio scale instead of the additive scale when the mediator is continuous. Albert and Nelson (2011) developed an approach for estimating path-specific effects in the context of a directed acyclic graph (DAG) using log linear models, and Albert 2012 considered an inverse-probability weighted estimator for the mediation effect on count outcomes.

For zero-inflated outcomes, different approaches (Min and Agresti 2002) have been proposed in the non-mediation context, including the zero-inflated Poisson (ZIP) (Lambert 1992) or zero-inflated Negative Binomial (ZINB) (Long 1997) model. Under the mediation context, the ZIP outcome distribution is:

$$P(Y_i^{Z_i, M_i^{Z_i}} = 0) = \omega_i + (1 - \omega_i)e^{-\lambda_i};$$
$$P(Y_i^{Z_i, M_i^{Z_i}} = j) = (1 - \omega_i)\frac{e^{-\lambda_i}\lambda_i^j}{j!}; \quad j > 0 \tag{10.8}$$

while the ZINB outcome distribution is:

$$P(Y_i^{Z_i, M_i^{Z_i}} = 0) = \omega_i + (1 - \omega_i)(1 + \sigma\lambda_i)^{-\frac{1}{\sigma}};$$
$$P(Y_i^{Z_i, M_i^{Z_i}} = j) = (1 - \omega_i)\frac{\Gamma(j + \frac{1}{\sigma})}{j!\Gamma(\frac{1}{\sigma})}(\sigma\lambda_i)^j(1 + \sigma\lambda_i)^{-j-\frac{1}{\sigma}}; \quad j > 0 \tag{10.9}$$

$$\text{where } \log\frac{\omega_i}{1 - \omega_i} = \alpha_{Y1} + \beta_{Y1}Z_i + \gamma_{Y1}M_i^{Z_i} + \xi_{Y1}Z_iM_i^{Z_i} + \eta_{Y1}^T X_i,$$
$$\log\lambda_i = \alpha_{Y2} + \beta_{Y2}Z_i + \gamma_{Y2}M_i^{Z_i} + \xi_{Y2}Z_iM_i^{Z_i} + \eta_{Y2}^T X_i \tag{10.10}$$
$$\sigma(\geq 0) \text{ is a dispersion parameter that does not depend on covariates.}$$

The basic idea of these zero-inflated models is that the outcome is a mixture of zeros and Poisson (or Negative Binomial) random variables with the mixture proportion $p(Z_i, M_i^{Z_i}, X_i)$ and Poisson (or Negative Binomial) mean $\lambda(Z_i, M_i^{Z_i}, X_i)$. Note that when an interpretation only relies on the second part (positive outcome) of the ZIP or ZINB model, the conclusion could be misleading because the two groups with the positive outcome are not ensured to be comparable by randomization (Follmann et al. 2009).

For count outcomes with a lot of zeros, more than expected under Poisson or Negative Binomial, Wang and Albert (2012) assumed a zero-inflated Negative Binomial (ZINB) model for the outcome, provided a mediation formula for the mediation effect estimation in a two-stage model, and then decomposed the mediation effect in a three-stage model when there is no post-treatment confounder.

Instead of decomposing the indirect (mediation) effect into different components for ZINB data (Wang and Albert 2012), Cheng et al. 2017 extended Imai et al.'s approach (2010b) for estimating the overall direct, indirect (mediation), and total effects specifically for zero-inflated count outcomes (zero-inflated Poisson or ZIP, and zero-inflated Negative Binomial or ZINB) in addition to count outcomes (Poisson and NB). Although the second part of a zero-inflated model alone may lead to misleading results, Cheng et al.'s approach (2017) does not only rely on the estimated coefficients in the second part of the ZIP or ZINB model but instead estimates the direct and indirect effects of treatment as a difference in corresponding potential zero-inflated count outcomes. Their approach uses information from all the randomized subjects with both parts of the model so that the ignorability of treatment holds. They adopt Imai et al.'s procedure (Imai et al. 2010a, b) based on the quasi-Bayesian Monte Carlo approximation of King et al. (2000). Specifically, the procedure (Imai et al. (2010a, b); Cheng et al. 2017a) involves multiple steps: (I) Fit the mediator and outcome models based on observed mediator (10.3) and outcome (10.8) or (10.9)

with (10.10), and obtain estimated model coefficients and their estimated asymptotic covariance matrix. (II) Simulate model coefficients from their sampling distribution based on the approximate multivariate normal distribution with mean and variance equal to the estimated coefficients and their estimated asymptotic covariance matrix obtained in (I), and sample K copies of the mediator and outcome model coefficients from their sampling distributions: θ_M^k and θ_Y^k. (III) For each copy $k = 1, ..., K$, (IIIa) simulate potential values of the mediator under each $z = 0, 1$ for each subject based on the mediator model (10.3) with simulated parameters (coefficients) obtained in (II); (IIIb) simulate potential outcomes under each $z = 0, 1$ for each subject based on the outcome model (10.4) with simulated potential mediator values obtained in (IIIa) and simulated parameters (coefficients) obtained in (II); (IIIc) compute the direct, mediation, and total treatment effects by averaging the difference between the corresponding two predicted potential outcomes discussed in Sect. 10.2. And (IV) compute the point estimates of direct, indirect (mediation), and total effects, confidence intervals and p values based on the results from J repetitions. The sample median, standard deviation, and percentiles of the corresponding distributions from the J repetitions are used as the point estimate, standard error, and confidence interval for the direct, indirect (mediation), and total effects.

10.5.1.2 Mediation Analysis with Post-baseline Confounders

Previously, we discussed causal mediation analysis for studies only with measured baseline confounders X_i. However, it is not uncommon that some confounding occurred after baseline. For example, in the DDHP MI-DVD study, when evaluating the direct effect of the MI+DVD intervention versus DVD alone on children's dental outcomes and its mediation effect through the mediator if whether or not caregivers made sure their child brushed at bedtime. Other intermediate variables (such as caregivers' oral hygiene knowledge and their own oral health-related behaviors) after baseline could be associated with both the mediator (whether or not they made sure their child brushed) and the outcome (children's dental outcomes), so those intermediate variables would be post-baseline confounders for the mediation analysis of interest.

Figure 10.3 shows the treatment mechanism through and around the mediator when the treatment (a) does not affect and (b) does affect the post-baseline confounder, respectively, where U_i denote post-baseline confounders.

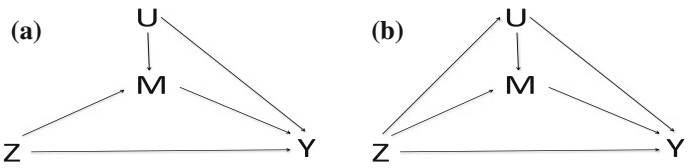


Fig. 10.3 Treatment mechanism when Z does not affect U (**a**) and when Z affects U **b**

Post-Baseline Confounders not Affected by the Treatment

Under the sequential ignorability (10.11), average natural effects are identified (Ten Have and Joffe 2010) when treatment Z_i does not affect the post-baseline confounder U_i (Fig. 10.3(a)):

$$(Y_i^{z^*,m}, M_i^z) \perp Z_i | X_i = x; \quad \text{and } Y_i^{z^*,m} \perp M_i^z | Z_i = z, X_i = x, U_i = u, \text{ for all } z, z^*, m, u. \tag{10.11}$$

Same as the first part of (10.1), the first part of (10.11) implies the ignorability of treatment; that is, the treatment is randomly assigned conditional on X_i. Although similar to the second part of (10.1) regarding the ignorability of the mediator, the second part of (10.11) is conditional on not only the treatment assignment and baseline covariates but also post-baseline confounders. Assumption (10.11) indicates that among subjects in the same treatment group who have the same values of baseline characteristics and post-baseline confounders, the mediator is effectively random (independent of confounding). In the MI+DVD study, the sequential ignorability assumes (1) the MI+DVD intervention is independent of confounders conditional on baseline covariates, and (2) the mediator (whether or not caregivers made sure their child brushed at bedtime) is independent of confounders conditional on assigned treatment group, baseline covariates, and post-baseline confounders such as caregivers' oral hygiene knowledge and behaviors.

To estimate the natural direct and indirect effects of the treatment when the post-baseline confounder U_i is not affected by treatment Z_i, the mediation model stays the same as (10.3), but the outcome model (10.4) needs to be modified by including the post-baseline confounder(s) in the model (Ten Have and Joffe 2010):

$$Y_i^{Z_i, M_i^{Z_i}} \sim f_Y \left(\theta_Y = g^{-1}(\alpha_Y + \beta_Y Z_i + \gamma_Y M_i^{Z_i} + \eta_Y^T X_i + \phi_Y^T U_i) \right) \tag{10.12}$$

For zero-inflated count data, the post-baseline confounder will be included in the outcome model as (10.10):

$$\begin{aligned} \log \frac{\omega_i}{1-\omega_i} &= \alpha_{Y1} + \beta_{Y1} Z_i + \gamma_{Y1} M_i^{Z_i} + \eta_{Y1}^T X_i + \phi_{Y1}^T U_i, \\ \log \lambda_i &= \alpha_{Y2} + \beta_{Y2} Z_i + \gamma_{Y2} M_i^{Z_i} + \eta_{Y2}^T X_i + \phi_{Y2}^T U_i \end{aligned}$$

Following Imai et al.'s general procedure (2010b), Cheng et al. (2017) discussed the method for estimating the natural direct and indirect (mediation) effects based on the mediation formula for count and zero-inflated count outcome when the post-baseline confounders are not affected by the treatment.

Post-Baseline Confounders Affected by the Treatment

When we evaluate the effect of the MI+DVD intervention on children's dental outcomes around or through the mediator whether or not caregivers made sure their child brushed at bedtime, we note that the MI+DVD intervention could also affect caregivers' oral hygiene knowledge and other related behaviors, which could be asso-

ciated with both the mediator whether or not they made sure their child brushed at bedtime and their child's dental outcomes. That is, there could be some post-baseline confounders on the mediator–outcome relationship, which are also affected by the treatment.

When treatment Z_i affects the post-baseline confounder U_i (Fig. 10.3(b)), the sequential ignorability (10.11) does not identify average natural effects without additional assumptions. Under (10.11) and the extended outcome model (10.13), although the estimate of the average controlled direct effect by $\hat{\beta}_Y$ could be biased, the average controlled mediation effect can be consistently estimated by a function of $\hat{\gamma}_Y$ (Ten Have and Joffe 2010):

$$Y_i^{Z_i,M_i} \sim f_Y(\theta_Y = g^{-1}(\alpha_Y + \beta_Y Z_i + \gamma_Y M_i + \eta_Y^T X_i + \phi^T U_i)). \tag{10.13}$$

The biased average controlled direct effect is due to the fact that the treatment Z_i affects the post-baseline confounder U_i and this treatment effect on U_i is not incorporated in the estimation of the controlled direct effect. For continuous outcomes with an identity link function in (10.13), Vansteelandt (2009) and Joffe and Greene (2009) used a two-stage ordinary least squares (OLS) procedure to estimate the average controlled direct effect by correcting the bias in the second stage. Their approaches involve two stages: (1) Fit an OLS model for outcome on randomization, mediator, and post-baseline confounder in the first stage $Y_i = \alpha_Y + \beta_Y Z_i + \gamma_Y M_i + \eta_Y^T X_i + \phi^T U_i$ and obtain $\hat{\gamma}_Y$; and (2) in the second stage, compute the adjusted outcome as $Y^{adj} = Y - \hat{\gamma}_Y M$, and then fit another OLS model for the adjusted outcome on randomized treatment as $Y^{adj} = \beta_Y^{adj} Z_i$. The estimated coefficient $\hat{\beta}_Y^{adj}$ will then be a good estimate for the average controlled direct effect when the outcome is continuous and there is a post-baseline confounder U_i affected by Z_i. However, for other types of outcomes such as count and zero-inflated count outcomes, the two-stage OLS procedure may not work well to adjust the bias in the estimation of the average controlled direct effect.

Alternatively, additional assumptions have been adopted for identification of the effect, sometimes in sub-populations or strata, when there are post-baseline confounders affected by treatment. Tchetgen Tchetgen and Shpitser (2012) and VanderWeele and Chiba (2014) considered a sensitivity analysis with various contrasts of the outcome between two sub-populations as sensitivity parameters and then corrected the bias with specified values of sensitivity parameters. Tchetgen Tchetgen and VanderWeele (2014) assumed monotonicity about the treatment effect on the post-baseline confounder and showed the nonparametric identifiability of the natural direct effect. When the mediator is binary, Taguri and Chiba (2015) classified subjects into four principal strata based on the joint distribution of the potential mediator under treatment and control and estimated the natural direct and indirect effects under an additional monotonicity assumption on treatment by-mediator effect and an assumption of common average mediator effects between compliant and never intermediates.

Instead of considering point identification of the effect, some researchers also considered the derivation of bounds for the natural direct and indirect effects, includ-

ing the work by Sjolander (2009), Kaufman et al. (2009) and Robins and Richardson (2011).

For count and zero-inflated count outcomes when there are post-baseline confounders affected by treatment, Cheng et al. (2017) discussed a series of assumptions, including Albert and Nelson's conditional independence assumption (Albert and Nelson 2011), for the identification of effects. Given the theoretical proofs on the identification under the series of assumptions, Cheng et al. (2017) proposed sensitivity analyses for count and zero-inflated count outcomes when there is post-baseline confounding. In (10.14), they define the mediation effect as the causal effect of the treatment on the outcome specifically through the mediator M under treatment z, and the direct effect as all other causal effects of the treatment on the outcome around M, including the effect through the post-baseline confounder U. That is, the confounding effect from U is included in the direct effect when it is not the interest. See Sect. 10.5.1.3 for discussion on mediation analyses when mediation effects via multiple mediators are of interest.

$$\begin{aligned} N\bar{M}E_z &= E\left(Y_i^{z,U_i^z,M_i^{1,U_i^1}} - Y_i^{z,U_i^z,M_i^{0,U_i^0}}\right), \quad \text{for } z = 0, 1 \\ N\bar{D}E_z &= E\left(Y_i^{1,U_i^1,M_i^{z,U_i^z}} - Y_i^{0,U_i^0,M_i^{z,U_i^z}}\right), \quad \text{for } z = 0, 1 \\ N\bar{T}E &= E\left(Y_i^{1,U_i^1,M_i^{1,U_i^1}} - Y_i^{0,U_i^0,M_i^{0,U_i^0}}\right) = N\bar{D}E_1 + N\bar{M}E_0. \end{aligned} \tag{10.14}$$

When there are post-baseline confounders affected by treatment, sequential ignorability (10.15) and (10.16), mediator models, and outcome models will be modified accordingly:

$$\left(Y_i^{z,u,m}, M_i^{z^*,u^*}, U_i^z\right) \perp Z_i \mid X_i = x \tag{10.15}$$

$$Y_i^{z,u,m} \perp M_i^{z^*,u^*} \mid X_i = x, Z_i = z, U_i^z = u \tag{10.16}$$

$$M_i^{Z_i} \sim f_M\left(\theta_M = h^{-1}(\alpha_M + \beta_M Z_i + \phi_M U_i^{Z_i} + \eta_M^T X_i)\right) \tag{10.17}$$

$$Y_i^{Z_i,M_i^{Z_i}} \sim f_Y\left(\theta_Y = g^{-1}(\alpha_Y + \beta_Y Z_i + \gamma_Y M_i^{Z_i} + \phi_Y U_i^{Z_i} + \eta_Y^T X_i)\right) \tag{10.18}$$

Additionally, Cheng et al. (2017) assumed various models for the post-baseline confounder $U_i^{Z_i}$ for the effect identification. The idea is that if we know the treatment mechanism on the post-baseline confounders, we are able to incorporate that information into the direct and mediation effect estimation. In a real-life study, although it is almost impossible to completely know such treatment mechanisms on confounders, investigators often understand partial information regarding such mechanisms based on their previous work and literature. Therefore, it is very helpful to have a sensitivity

analysis incorporating such information and see how the direct and mediation effects will change when specific values of the parameters are changed under the uncertainty in the treatment mechanisms on post-baseline confounders. Cheng et al. (2017) considered various models (10.19)–(10.21) for continuous post-baseline confounders U, where model (10.21) allows the heterogeneity treatment effect on U for individuals. For a binary confounder U, one can assume that there is an underlying continuous variable under U, which follows one of the following models (10.19)–(10.21).

$$U_i^1 = U_i^0 + \beta_U, \tag{10.19}$$

$$U_i^1 = U_i^0 + \beta_U + \tau_U^T X_i, \tag{10.20}$$

$$U_i^1 = U_i^0 + \beta_U + \tau_U^T X_i + \delta_i, \text{ where } \delta_i \perp (Z_i, X_i, U_i^0, Y_i^{z,u,m}, M_i^{z*,u*}) \text{ and } \delta_i \text{ follows a known distribution} \tag{10.21}$$

For general post-baseline confounders, Cheng et al. (2017) considered a set of assumptions to identify the effects (10.14),

$$\left(Y_i^{z,u,m}, M_i^{z^*,u^*}, U_i^1, U_i^0\right) \perp Z_i \mid X_i = x \tag{10.22}$$

$$Y_i^{z,u,m} \perp M_i^{z^*,u^*} \mid X_i = x, Z_i = z, U_i^0 = u, U_i^1 = u^* \tag{10.23}$$

and

$$U_i^{Z_i} \sim f_U\left(\theta_U = o^{-1}(\alpha_U + \beta_U Z_i + \tau_U^T X_i)\right),$$
$$U_i^1 \perp U_i^0 \mid X_i = x \text{ and } \left(Y_i^{z,u,m}, M_i^{z^*,u^*}\right) \perp \left(U_i^0, U_i^1\right) \mid X_i = x, Z_i = z \tag{10.24}$$

Note that (10.22) and (10.23) involve the joint distribution of $\left(U_i^0, U_i^1\right)$ while (10.15) and (10.16) involve the marginal distribution U_i^z. Therefore, (10.22) and (10.23) are stronger than (10.15) and (10.16). Assumption (10.24) is similar the conditional independence assumption on $Z_1(1)$ and $Z_0(0)$ in Albert and Nelson (2011), but (10.24) assumes some relation between U^1 and U^0 instead of assuming independence between U^1 and U^0 as in Albert and Nelson (2011) and hence could be more practical in some real-life studies.

Cheng et al. (2017) showed that the average effects $N\bar{M}E_z$, $N\bar{D}E_z$ and $N\bar{T}E_z$ can be identified under (A) sequential ignorability (10.15) and (10.16), mediator model (10.17) and outcome model (10.18), and one of confounder models (10.19)–(10.21); or (B) sequential ignorability (10.16), (10.22) and (10.23), mediator model (10.17) and outcome model (10.18), and the confounder model (10.24). The same results will follow when the treatment by-confounder interaction $Z_i \times U_i^{Z_i}$ is included in the mediator model (10.17) and both the treatment by-mediator and treatment by-confounder interactions $Z_i \times M_i^{Z_i}$ and $Z_i \times U_i^{Z_i}$ are included in the outcome model (10.18). The procedure based on the quasi-Bayesian Monte Carlo approximation (King et al. 2000) discussed in Sect. 10.5.1.1 can still be used for inference on the

direct, mediation, and total treatment effects, except that one additional confounder model (10.19), (10.20), (10.21) or (10.24) will be incorporated.

As discussed above, in a real study, we are not able to know the values of parameters β_U and τ_U^T for sure. However, the study can provide a reasonable range on the potential values along with information from literature. Then we can vary the values of parameters β_U and τ_U^T one or two at a time and see how the estimates of effects (10.14) will change for a sensitivity analysis. Cheng et al. (2017) proposed to use estimates from a regression of observed U_i on treatment Z_i, covariates X_i and their interaction as reasonable starting points for the choice of values for β_U and τ_U^T in the sensitivity analysis. For example, in $U_i = \alpha_U + \delta_U Z_i + \nu_U X_i + \epsilon_i$, use $\hat{\delta}_U \, pmc\%$ as a range for the parameters, where the choice of $c\%$ (say $\frac{1}{3}$, $\frac{1}{2}$ or 1 of the estimate) will be based on expert knowledge in a study to represent the possible treatment effect on the confounder. Then 10–20 equally divided values in the range can be used for the sensitivity analysis.

In the MI+DVD study, the MI+DVD intervention might change the times children visited their dentists during the follow-up, and the dental visits were possibly associated with both the mediator (whether or not parents made sure children brushed their teeth) and children's dental outcomes. To account for the post-baseline dental visits, a sensitivity analysis is helpful to understand how the effects would change with varying post-treatment confounder effect. Because the MI+DVD intervention had a small effect in reducing dental visits based on observed data (-0.15), $-0.15 \pm \frac{1}{3}(-0.15)$; i.e., $(-0.20, -0.10)$ was used as the reasonable range for β_U in terms of possible intervention–confounder effect. Figure 10.4 shows that with various values of β_U, the mediation effects stay around 0 while the direct and total effects increase and vary within a range from 0.03 for untreated decayed surfaces (cavities).

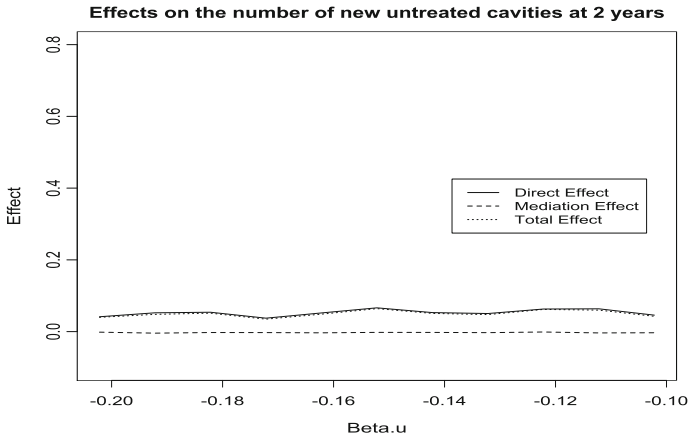


Fig. 10.4 Sensitivity analysis for direct, mediation, and total effects on the numbers of new untreated decayed tooth surfaces (cavities) with varying treatment effects on the post-treatment confounder β_U

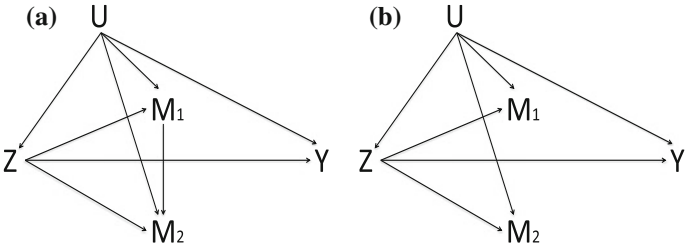


Fig. 10.5 Causal diagram with treatment Z, mediators M_1 and M_2, outcome Y, and confounders U under **a** M_1 causally affects M_2 and **b** M_1 does not causally affect M_2

10.5.1.3 Mediation Analysis with Multiple Mediators

In real studies, the treatment often affects multiple intermediate variables. Figure 10.5 shows a causal diagram with two mediators. Daniel et al. (2015) discussed various approaches when more than one intermediate variables exist in a study. The existing approaches can be grouped into three types (1) M_2 is the mediator of interest, and M_1 is treated as a mediator–outcome confounder affected by treatment (Vansteelandt and VanderWeele 2012; Tchetgen and VanderWeele 2014; VanderWeele and Chiba 2014; VanderWeele et al. 2014; Taguri and Chiba 2015; Cheng et al. 2017); (2) path-specific effects are estimated, but their sum does not equal the total effect (Avin et al. 2005; Albert and Nelson 2011); and (3) the multiple mediators do not causally affect one another (MacKinnon 2000; Preacher and Hayes 2008; Lange et al. 2014; Taguri et al. 2018), see Fig. 10.5b.

In Sect. 10.5.1.2, we have discussed the first type of mediation analyses when mediation from a specific intermediate variable is of interest. Then other intermediate variables are considered as post-baseline confounders and included in the direct effect of the treatment. However, in some studies, investigators are interested in the mediation effects via multiple mediators. Imai and Yamamoto (2013) assumed a linear structural equation model for the outcome and mediators and estimated the effects; Daniel et al. (2015) decomposed the total effect in the finest possible way, and VanderWeele and Vansteelandt (2013) considered the mediators one at a time as joint mediators and proposed decomposition of the total effect with regression-based and weighting approaches.

For count data, Albert and Nelson (2011) assumed independence between one mediator under treatment $Z_1(1)$ and under control $Z_1(0)$ and then conducted a sensitivity analysis on path-specific effects, where the sum of path-specific effects does not equal the total effect. Taguri et al. (2018) considered the setting with causally non-ordered multiple mediators. It has become common to have a treatment with multiple components targeting on multiple non-causally related mediators. For example, a common cavity management strategy includes two components—an antibacterial component to reduce oral bacteria and a mineralization component to strengthen the teeth with fluoride. In such studies, the mediators are often not causally related. The sequential ignorability becomes

$$
\begin{aligned}
Y^{z,m_1,m_2} &\coprod Z|U \text{ for all } (z, m_1, m_2),\\
Y^{z,m_1,m_2} &\coprod (M_1^z, M_2^z)|U, Z = z \text{ for all } (z, m_1, m_2),\\
M_1^z, M_2^z &\coprod Z|U \text{ for all } z,\\
Y^{z,m_1,m_2} &\coprod (M_1^{z^*}, M_2^{z^*})|U,\\
M_1^{z^*} &\coprod M_2^{z^{**}}|U, \text{ for all } (z, z^*, z^{**}, m_1, m_2).
\end{aligned}
$$

Taguri et al. (2018) decomposed the total effect as a function of the indirect (mediation) effects through multiple causally non-ordered mediators and the direct effect around the mediators:

$$
\begin{aligned}
Y_i^1 - Y_i^0 &= (Y_i^{1,M_1^1,M_2^1} - Y_i^{1,M_1^0,M_2^0}) + (Y_i^{1,M_1^0,M_2^0} - Y_i^{0,M_1^0,M_2^0})\\
&= \text{total natural indirect effect} + \text{pure natural direct effect}\\
&= (Y_i^{1,M_1^1,M_2^1} - Y_i^{0,M_1^1,M_2^1}) + (Y_i^{0,M_1^1,M_2^1} - Y_i^{0,M_1^0,M_2^0})\\
&= \text{total natural direct effect} + \text{pure natural indirect effect}
\end{aligned}
$$

where the "pure" and "total" effects capture the differential inclusion of the interaction between the treatment and mediators.

Taguri et al. (2018) also provided an analytical approach for the joint natural indirect effect between two mediators as a function of the indirect effect for individual mediator and the mediated interactive effect:

$$
\begin{aligned}
Y_i^{1,M_1^1,M_2^1} - Y_i^{1,M_1^0,M_2^0} &= (Y_i^{1,M_1^1,M_2^0} - Y_i^{1,M_1^0,M_2^0}) + (Y_i^{1,M_1^0,M_2^1} - Y_i^{1,M_1^0,M_2^0}) \quad (10.25)\\
&+ (Y_i^{1,M_1^1,M_2^1} - Y_i^{1,M_1^1,M_2^0}) + (Y_i^{1,M_1^0,M_2^1} - Y_i^{1,M_1^0,M_2^0}) \quad (10.26)\\
&= PSE_1^0 + PSE_2^0 + MInt, \quad (10.27)
\end{aligned}
$$

where $PSE_1^0 = (Y_i^{1,M_1^1,M_2^0} - Y_i^{1,M_1^0,M_2^0})$ is the indirect effect through M_1 while M_2 takes the value under control, $PSE_2^0 = (Y_i^{1,M_1^0,M_2^1} - Y_i^{1,M_1^0,M_2^0})$ is the indirect effect through M_2 while M_1 takes the value under control, and MInt is the mediated interaction between M_1 and M_2. By (10.25), we see that even if the two mediators are not causally ordered, the sum of two indirect effects considered separately may not be the same as the joint indirect effect when there are interactions between the effects through the two mediators. When MInt=0, then the joint total natural indirect effect is the same as the sum of the two separate total indirect effect. Similarly the population average effect can be decomposed as the decomposition (10.25) at the individual level. Understanding how the treatment works through the multiple mediators, relative contributions of different components and their interactions provide additional information to design a better treatment strategy.

10.5.2 Methods Without the Assumption of Sequential Ignorability

In Sect. 10.5.1, we discussed various mediation methods assuming sequential ignorability for count and zero-inflated count data (Albert and Nelson 2011; Wang and Albert 2012; Albert 2012; Taguri et al. 2018; Cheng et al. 2017). Sequential ignorability is plausible for some studies. However, the sequential ignorability may not be plausible for other studies. For example, in DDHP MI+DVD study, caregivers' and children's oral health-related behaviors were not randomized or controlled by the investigators but could be affected by factors other than the MI+DVD intervention, such as oral health education the parents/children received from family dentists, schools, communities, or the Internet. Those outside factors were not measured in the study and could be confounders of the mediator–outcome relationship for being associated with whether or not the parents made sure children brushed their teeth and children's dental outcome, so the sequential ignorability may not hold in this study.

When unmeasured confounding is of concern in a study, instrumental variable (IV) methods are very helpful for obtaining consistent estimates for treatment effects by adjusting for both unmeasured and measured confounders when a valid and strong IV can be found (Angrist et al. 1996). In the context of mediation analysis, a valid IV is a variable that, given the measured baseline variables: (1) affects the value of the mediator; (2) is independent of the unmeasured confounders; and (3) does not have a direct effect on the outcome other than through its effect on the mediator. Methods for mediation analysis based on the IV approach have been proposed by investigators (Ten Have et al. 2007; Albert 2008; Dunn and Bentall 2007; Small 2012) using the randomization interacted with baseline covariates as IVs, but those methods focus on linear models for continuous outcomes. When a linear model holds for the outcome, two-stage least squares (2SLS) provides consistent estimates for causal effects when there is a valid IV. When the outcome model is nonlinear, two-stage residual inclusion (2SRI) and two-stage predictor substituion (2SPS) (Nagelkerke et al. 2000; Terza et al. 2008) have been proposed to estimate causal effects in a general context. See Guo and Small (2016) and Cai et al. (2011) for comparison of 2SRI and 2SPS in general settings. In the context of mediation analysis when there is a concern of unmeasured confounding such that the assumption of sequential ignorability for the mediator might fail, Guo et al. (2018) developed a new IV approach using the randomization by-baseline covariate interaction $Z \times \mathbf{X}^{\mathrm{IV}}$ as the instrumental variable for count and zero-inflated count data. Since the randomized treatment itself is not used as the IV, both the direct and indirect effects of the treatment on the outcome of interest can be estimated. They considered the controlled and natural direct and indirect effects on a ratio scale for count and zero-inflated count.

$$\text{Controlled effect ratio:} \quad \text{direct}\left(z \text{ vs. } z^*; m, \mathbf{x}, u\right) \quad \frac{E\left(Y\left(z, m\right)|\mathbf{x}, u\right)}{E\left(Y\left(z^*, m\right)|\mathbf{x}, u\right)}, \tag{10.28}$$

$$\text{indirect}\left(m \text{ vs. } m^*; z, \mathbf{x}, u\right) \quad \frac{E\left(Y\left(z, m\right)|\mathbf{x}, u\right)}{E\left(Y\left(z, m^*\right)|\mathbf{x}, u\right)}; \tag{10.29}$$

$$\text{Natural effect ratio:} \quad \text{direct}\left(z \text{ vs. } z^*; M^{z^*}, \mathbf{x}, u\right) \quad \frac{E\left(Y\left(z, M^{z^*}\right)|\mathbf{x}, u\right)}{E\left(Y\left(z^*, M^{z^*}\right)|\mathbf{x}, u\right)}, \tag{10.30}$$

$$\text{indirect}\left(M^z \text{ vs. } M^{z^*}; z, \mathbf{x}, u\right) \quad \frac{E\left(Y\left(z, M^z\right)|\mathbf{x}, u\right)}{E\left(Y\left(z, M^{z^*}\right)|\mathbf{x}, u\right)}; \tag{10.31}$$

Considering a generalized linear outcome model with log link for count and a Neyman Type A distributed outcome (Dobbie and Welsh 2001) for zero-inflated count:

$$f\{E\left(Y(z, m)|\mathbf{x}, u\right)\} = \beta_0 + \beta_z z + \beta_m m + \beta_{\mathbf{x}}\mathbf{x} + u, \tag{10.32}$$

where u is an unmeasured confounder, Guo et al. (2018) showed that estimating the controlled effect ratios is equivalent to estimating β_z and β_m

$$\frac{E\left(Y\left(z, m\right)|\mathbf{x}, u\right)}{E\left(Y\left(z^*, m\right)|\mathbf{x}, u\right)} = \exp\left(\beta_z(z - z^*)\right), \quad \frac{E\left(Y\left(z, m\right)|\mathbf{x}, u\right)}{E\left(Y\left(z, m^*\right)|\mathbf{x}, u\right)} = \exp\left(\beta_m(m - m^*)\right). \tag{10.33}$$

and that the natural direct effect ratio will be the same as the controlled direct effect ratio:

$$\frac{E\left(Y\left(z, M^{z^*}\right)|\mathbf{x}, u\right)}{E\left(Y\left(z^*, M^{z^*}\right)|\mathbf{x}, u\right)} = \exp\left(\beta_z(z - z^*)\right). \tag{10.34}$$

However, given the mediator model (10.35) and outcome model (10.36), the natural indirect effect ratio will be different from the controlled indirect effect ratio, being (10.37) for a continuous mediator and (10.38) for a binary mediator:

$$h\left(E\left(M^{z^*}|\mathbf{x}, u\right)\right) = \alpha_0 + \alpha_z z^* + \alpha_{\mathbf{x}}\mathbf{x} + \alpha_{\text{IV}} z^* \mathbf{x}^{\text{IV}} + u, \tag{10.35}$$

$$f\{E\left(Y(z, M^{z^*})|\mathbf{x}, u\right)\} = \beta_0 + \beta_z z + \beta_m M^{z^*} + \beta_{\mathbf{x}}\mathbf{x} + u, \tag{10.36}$$

$$\frac{E\left(Y\left(z, M^z\right)|\mathbf{x}, u\right)}{E\left(Y\left(z, M^{z^*}\right)|\mathbf{x}, u\right)} = \exp\left(\beta_m\alpha_z\left(z - z^*\right) + \beta_m\alpha_{\text{IV}}\mathbf{x}^{\text{IV}}(z - z^*)\right), \quad \text{continuous} \tag{10.37}$$

$$= \frac{P(M^z = 1|\mathbf{x}, u)\exp(\beta_m) + P(M^z = 0|\mathbf{x}, u)}{P(M^{z^*} = 1|\mathbf{x}, u)\exp(\beta_m) + P(M^{z^*} = 0|\mathbf{x}, u)}, \quad \text{binary.} \tag{10.38}$$

For a continuous mediator, the natural direct effect ratio (10.34) and indirect effect ratio (10.37) are identifiable given that the parameters β_z, β_m, α_z and α_{IV} can be estimated consistently. However, the natural indirect effect ratio for a binary mediator (10.38) depends on the values of the unmeasured u and therefore is not identifiable without additional assumptions.

Using the randomization by-baseline covariate interaction $Z \times \mathbf{X}^{\text{IV}}$ as an IV, both 2SRI and 2SPS fit the same mediator model in Stage I based on observables: $h\{E\left(M|z, \mathbf{x}, z \times \mathbf{x}^{\text{IV}}\right)\} = \alpha_0 + \alpha_z z + \alpha_{\text{IV}} z \times \mathbf{x}^{\text{IV}} + \alpha_{\mathbf{x}}\mathbf{x}$. In Stage II, 2SRI fits the

outcome model with the residual $\hat{r} = (m - \hat{m})$ from Stage I : $f\{E\left(Y|m, z, \mathbf{x}, \hat{r}\right)\} = \beta_0 + \beta_z z + \beta_m m + \beta_{\mathbf{x}}\mathbf{x} + \beta_r \hat{r}$; 2SPS uses the predicted value $\hat{m}$ instead of the residual $\hat{r}$ from the first stage for the outcome model $f\{E\left(Y|\hat{m}, z, \mathbf{x}\right)\} = \beta_0 + \beta_z z + \beta_m \hat{m} + \beta_{\mathbf{x}}\mathbf{x}$. Guo et al. (2018) showed that when the outcome model is log linear and the mediator model is linear, 2SRI provides consistent estimate of the parameters but 2SPS is not necessarily consistent; when the outcome model is log linear and the mediator model is logit, neither 2SRI nor 2SPS is consistent.

Because the two-stage methods may not be consistent for a count or zero-inflated count especially when the mediator is not continuous, Guo et al. (2018) considered a new approach using a set of estimating functions to incorporate information about the parameters and distribution such that $E\{g(w, \theta)\} = 0$, where $w = (z, \mathbf{x}, m, y)$ and $\theta = (\beta_0, \beta_z, \beta_m, \beta_{\mathbf{x}})$ are the parameters associated with the outcome model. Then they maximized the empirical likelihood $\prod_{i=1}^{n} p_i$ subject to the restrictions $p_i \geq 0$, $\sum_{i=1}^{n} p_i = 1$, $\sum_{i=1}^{n} p_i g(w_i, \theta) = 0$ for $\hat{\theta}$, where p_i is the probability to observe the data (Z_i, X_i, M_i, Y_i). Guo et al. (2018) showed that their estimating equations and empirical likelihood estimator $\hat{\theta}$ is consistent under some mild regularity conditions. Note that this approach does not require the specification of the error distribution of the outcome and therefore is robust to the misspecification of the outcome distribution.

In the MI+DVD study, more than 60% of children had no new untreated lesions, so the count outcome has many zeros. The mediator of whether or not caregivers made sure their child brushed at bedtime is binary. There is a concern about unmeasured confounders of oral health education that the caregivers and/or children received from their dentists, schools, communities, or the Internet outside of the study. Therefore, the IV mediation analysis was considered. Three baseline covariates (the number of times that the child brushed at baseline, whether or not caregivers made sure their child brushed at bedtime at baseline, and whether or not caregivers provided the child healthy meals at baseline) are considered as important behavioral variables related to subsequent oral hygiene behaviors and oral health. Therefore, their interactions with the randomized intervention were used to construct three IVs. The three constructed IVs were shown to be reasonable IVs following a series of assessments on the IV assumptions (Guo et al. 2018). The result shows a controlled direct effect ratio of 1.081, indicating that the intervention did not have much direct effect on the number of new untreated decayed tooth surfaces (cavities). Parent behavior in making sure their child brushed at bedtime tended to decrease the number of new untreated decayed tooth surfaces (cavities) (controlled indirect effect ratio of 0.595) but the effect was not statistically significant with a 90% CI of (0.070, 4.604).

10.6 Summary

Mediation analysis provides helpful information for understanding mechanisms underlying risk factors or treatment on health outcomes and is a powerful tool for better-designed interventions. When linear models do not fit non-continuous health

outcomes, mediation analysis becomes challenging. This chapter has discussed conventional methods as well as recently developed causal methods for mediation analyses for binary, count, and zero-inflated count health outcomes under different settings and assumptions. Researchers are encouraged to evaluate the plausibility of assumptions and select a method appropriate for their study.

Acknowledgements The authors thank Dean Amid Ismail and Sungwoo Lim for providing the DDHP MI-DVD data which was performed with support from cooperative agreement U54 DE014261. This chapter was made possible by cooperative agreement U54 DE019285 from the National Institute of Dental and Craniofacial Research, a component of the United States National Institutes of Health.

References

Albert, J. M. (2008). Mediation Analysis via potential outcomes models. *Statistics in Medicine*, *27*(8), 1282–1304.

Albert, J. M., & Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics*, *67*(3), 1028–1038.

Albert, J. M. (2012). Mediation analysis for nonlinear models with confounding. *Epidemiology*, *23*, 879–888.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Avin, C., Shpitser, I., & Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 357–363). Edinburgh, UK: Morgan-Kaufmann.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Cai, B., Small, D., & Ten Have, T. (2011). Two-stage instrumental variable methods for estimating the causal odds ratio: Analysis of bias. *Statistics in Medicine*, *30*(15), 1809–1824.

Cheng, J., Cheng, N. F., Guo, Z., Gregorich, S. E., Ismail, A. I., & Gansky, S. A. (2017). Mediation analysis for count and zero-inflated count data. *Statistical Methods in Medical Research*. https://doi.org/10.1177/0962280216686131.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediation models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, *112*(4), 558–577.

Coxe, S., & MacKinnon, D. (2010). Mediation analysis of Poisson distributed count outcome. *Multivariate Behavioral Research*, *45*(6), 1022.

Daniels, M. J., Roy, J., Kim, C., Hogan, J. W., & Perri, M. G. (2012). Bayesian Inference for the Causal Effect of Mediation. *Biometrics*, *68*(4), 1028–1036.

Daniel, R. M., De Stavola, B. L., Cousens, S. N., & Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, *71*(1), 114.

Dobbie, M. J., & Welsh, A. H. (2001). Models for zero-inflated count data using the Neyman type A distribution. *Statistical Modeling*, *1*(11), 65–80.

Dunn, G., & Bentall, R. (2007). Modeling treatment effect heterogeneity in randomised controlled trials of complex interventions (psychological treatments). *Statistics in Medicine*, *26*, 4719–4745.

Elliott, M. R., Raghunathan, T. E., & Li, Y. (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics*, *11*(2), 353–372.

Follmann, D., Fay, M. P., & Proschan, M. (2009). Chop-Lump tests for vaccine trials. *Biometrics*, *65*(3), 885–893.

Frangakis, C.E., & Rubin, D.B. (2002). Principal stratification in causal inference. *Biometrics*, *58*, 21–29.

Gallop, R., Small, D. S., Lin, J., Elliott, M., Joffe, M., & Ten Have, T. (2009). Mediation analysis with principal stratification. *Statistics in Medicine*, *28*(7), 1108–1130.

Goetgeluk, S., Vansteelandt, S., & Goetghebeur, E. (2008). Estimation of controlled direct effects. *Journal of the Royal Statistical Society, Series B*, *70*(5), 1049–1066.

Guo, Z., Small, D. S., Gansky, S.A., & Cheng, J. (2018). Mediation analysis for count and zero-inflated count data without sequential ignorability and its application in dental studies. *Journal of the Royal Statistical Society, Series C*, *67*, 371–394.

Guo, Z., & Small, D. S. (2016). Control function instrumental variable estimation of nonlinear causal effect models. *Journal of Machine Learning Research*, *17*(1), 1–35.

Hafeman, D. M. (2009). 'Proportion Explained': A causal interpretation for standard measures of indirect effect? *American Journal of Epidemiology*, *170*(11), 1443–1448.

Holmbeck, G. N. (2002). Post-hoc probing of significant moderational and mediational effects in studies of pediatric populations. *Journal of Pediatric Psychology*, *27*(1), 87–96.

Imai, K., Keele, L., & Yamamoto, T. (2010a). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1), 51–71.

Imai, K., Keele, L., & Tingley, D. (2010b). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334.

Imai, K., & Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, *21*(2), 141–171.

Ismail, A. I., Ondersma, S., Willem Jedele, J. M., Little, R. J., & Lepkowski, J. M. (2011). Evaluation of a brief tailored motivational intervention to prevent early childhood caries. *Community Dentistry and Oral Epidemiology*, *39*(5), 433–448.

Jo, B. (2008). Causal inference in randomized experiments with mediational process. *Psychological Methods*, *13*(4), 314–336.

Jo, B., Stuart, E. A., MacKinnon, D. P., & Vinokur, A. D. (2011). The use of propensity scores in mediation analysis. *Multivariate Behavioral Research*, *46*(3), 425–452.

Joffe, M., & Greene, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics*, *65*(2), 530–538.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating medication in treatment evaluations. *Evaluation Review*, *5*(5), 602–619.

Kaufman, S., Kaufman, J. S., & MacLehose, R. F. (2009). Analytic bounds on causal risk differences in directed acyclic graphs involving three observed binary variables. *Journal of Statistical Planning and Inference*, *139*(10), 34733487.

King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science*, *44*(2), 341–355.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14.

Lange, T., Vansteelandt, S., & Bekaert, M. (2012). A simple unified approach for estimating natural direct and indirect effects. *American Journal of Epidemiology*, *176*(3), 190–195.

Lange, T., Rasmussen, M., & Thygesen, L. C. (2014). Assesing natural direct and indirect effects through multiple pathways. *American Journal of Epidemiology*, *179*(4), 513–518.

Loeys, T., Moerkerke, B., De Smet, O., Buysse, A., Steen, J., & Vansteelandt, S. (2013). Flexible mediation analysis in the presence of nonlinear relations: Beyond the mediation formula. *Multivariate Behavioral Research*, *48*, 871–894.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, *17*, 144–158.

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of effect measures. *Multivariate Behavioral Research*, *30*, 41–62.

MacKinnon, D. P. (2000). Contrast in multiple mediator models. In *Multivariate applications in substance use research* (pp. 141–160). Mahwah, NJ: Lawrence Erlbaum, Associates Publishers.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*, 83–104.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Erlbaum.

MacKinnon, D. P., & Luecen, L. J. (2011). Statistical analysis for identifying mediating variables in public health dentistry interventions. *Journal of Public Health Dentistry*, *71*(Suppl 1), S37–46.

Min, Y., & Agresti, A. (2002). Modeling nonnegative data with clumping at zero: a survey. *Journal of the Iranian Statistical Society*, *1*, 7–33.

Nagelkerke, N., Fidler, V., Bernsen, R., & Borgdorff, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine*, *19*, 1849–1864.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles (with discussion). Section 9 (translated). *Statistical Science*, *5*, 465–480.

Pearl, J. (2001). Direct and indirect effects. In J. Breese & D. Koller (Eds.), *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2012). The mediation formula: A guide to the assessment of causal pathways in nonlinear models. In C. Berzuini, P. Dawid, & L. Bernadinelli (Eds.), *Causality: Statistical perspectives and applications* (pp. 151–175). West Sussex, UK: Wiley.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavioral Research Methods*, *40*, 879–891.

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*, 143–155.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In P. J. Green, N. L. Hjort, & S. Richardson (Eds.), *Highly structured stochastic systems*. New York, NY: Oxford University Press.

Robins, J.M., & Richardson, T.S. (2011). Alternative graphical causal models and the identification of direct effects. In: Shrout P, (Ed.), *Causality and psychopathology: Finding the determinants of disorders and their cures*. Oxford University Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, *31*, 161–170.

Sjlander, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine*, *28*, 558571.

Small, D. S. (2012). Mediation analysis without sequential ignorability: Using baseline covariates interacted with random assignment as instrumental variables. *Journal of Statistical Research*, *46*, 91–103.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological methodology* (pp. 290–312). San Francisco, CA: Jossey-Bass.

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, *33*, 230–251.

Steyer, R., Mayer, A., & Fiege, C. (2014). Causal inference on total, direct, and indirect effects. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 606–631). Dordrecht, The Netherlands: Springer.

Taguri, M., & Chiba, Y. (2015). A principal stratification approach for evaluating natural direct and indirect effects in the presence of treatment-induced intermediate confounding. *Statistics in Medicine*, *34*, 131144.

Taguri, M., Featherstone, J., & Cheng, J. (2018). Causal mediation analysis with multiple causally non-ordered mediators. *Statistical Methods in Medical Research*, *27*, 3–19. https://doi.org/10.1177/0962280215615899.

Tchetgen Tchetgen, E. J., & Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, *40*, 18161845.

Tchetgen Tchetgen, E. J., & VanderWeele, T. J. (2014). Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*, *25*, 282291.

Ten Have, T. R., Joffe, M., Lynch, K., Maisto, S., Brown, G., & Beck, A. (2007). Causal mediation analyses with rank preserving models. *Biometrics*, *63*, 926–934.

Ten Have, T. R., & Joffe, M. (2010). A review of causal estimation of effects in mediation analyses. *Statistical Methods in Medical Research*. in press.

Terza, J., Basu, A., & Rathouz, P. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Health Economics*, *27*, 527–543.

Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, *18*, 137–150.

Van der Laan, M., & Petersen, M. (2008). Direct effect models. *International Journal of Biostatistics*, *4*, Article 23.

VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistical Probability Letter*.

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics in Its Interface*, *2*, 457–468.

VanderWeele, T. J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, *172*, 1339–1348.

VanderWeele, T. J., & Vansteelandt, S. (2013). Mediation analysis with multiple mediators. *Epidemiologic Methods*, *2*, 95–115.

VanderWeele, T. J., & Chiba, Y. (2014). Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiology, Biostatistics, and Public Health*, *11*, e9027.

VanderWeele, T. J., Vansteelandt, S., & Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, *25*, 300–306.

Vansteelandt, S. (2009). Estimating direct effects in cohort and case-control studies. *Epidemiology*, *20*, 851–860.

Vansteelandt, S. (2010). Estimation of controlled direct effects on a dichotomous outcome using logistic structural direct effect models. *Biometrika*, *97*, 921–934.

Vansteelandt, S., Bekaert, M., & Lange, T. (2012). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods*, *1*, 131–158.

Vansteelandt, S., & VanderWeele, T. J. (2012). Natural direct and indirect effects on the exposed: Effect demoposition under weak assumptions. *Biometrics*, *68*, 1019–1027.

Wang, W., & Albert, J. M. (2012). Estimation of mediation effects for zero-inflated regression models. *Statistics in Medicine*, *31*, 3118–3132.

Wright, S. (1920). The relative importance of heredity and environment in determing the piebald pattern of guinea pigs. *Proceedings of the National Academy of Science*, *6*, 320–332.

Chapter 11
Statistical Considerations for Quantitative Imaging Measures in Clinical Trials

Ying Lu

11.1 Introduction

Since the discovery of X-ray in 1895, noninvasive medical imaging techniques have experienced many major milestones, including inventions of computer tomography (CT), MRI, PET, and ultrasound; each of them revolutionized the medicine (Thomas and Banerjee 2013). Medical imaging has been extensively used in modern medicine for population screening and risk assessment, disease diagnosis, prognostic prediction, monitoring disease progress and treatment response, imaging-guided treatments, and as tools for basic biology. Another important application of imaging techniques is in drug developments. Imaging endpoints have been used in clinical trial to assess the effects of drugs on patients, such as primary or secondary clinical endpoints, or as surrogate markers for clinical events that may not be easily observed during the trials.

The first published randomized controlled clinical trial conducted by BMRC in 1946 studied the effect of streptomycin versus bed resting for treating pulmonary tuberculosis (No author 1948). X-ray films were used to assess patient eligibility as well as treatment effects. Two radiologists and one clinician performed independent and blinded readings to assess changes from baseline. Disagreement was resolved via consensus readings. Based on the radiologists' assessment, the trial concluded that streptomycin was "no doubt" more effective than bed resting.

Since then, imaging has been an integrated part of clinical trials. We have used number and size of lesions to evaluate the treatment effect for multiple sclerosis drugs, bone mineral density (BMD) and vertebral fractures for osteoporosis drugs, and the "response evaluation criteria in solid tumors" (RECIST) for cancer treatments. All of them are measured via imaging techniques.

Y. Lu (✉)
Department of Biomedical Data Science, Stanford University School of Medicine, HRP Redwood Building, T101B, Stanford, CA 94305-5405, USA
e-mail: ylu1@stanford.edu

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_11

The advantages of imaging techniques in clinical trials are many folds: They are noninvasive, can be measured repeatedly, are objective and reproducible measures of clinical changes, and are oftentimes measured both qualitatively and quantitatively. Their disadvantages include equipment and reader dependence, complex logistics, and higher costs.

While image is a form of data, direct overlap of medical images is not common used to compare the trial results. Radiologists have been used to assess the images in categorical values. More often, regions of interests (ROIs) were pre-defined and some summary statistics from the ROIs are derived as quantitative endpoints. As a result, there are also multiple ways to summarizing an image and selection of what is the most clinically relevant summary will require clinical insights of clinicians. However, many other properties, such as measurement errors of the summary statistics, predictive accuracy, and leading time for clinical events, are also important considerations (Lu and Zhao 2015). Thus, the selection of proper imaging endpoints requires collaborative efforts of a multidisciplinary team (Zhao 2010).

Clinical trials are experiments or observations done in clinical research. The clinical trials to validate imaging endpoints are different from the trials that using established imaging techniques to determine the effectiveness of medical intervention strategies. The American College of Radiology Imaging Network (ACRIN) has a consortium to evaluate medical imaging techniques (Hillman 2005). For this chapter, we will focus on the use of imaging techniques in clinical trials for therapeutic interventions. However, the methods are also relevant for general use of all kinds of biomarkers.

In the remaining of chapter, two topics are specifically discussed. Section 11.2 discusses statistical considerations of selecting an (imaging) biomarker endpoint for a phase II oncology trial. Section 11.3 discusses quality control and quality assurance of bone mineral density and contents in pediatric bone trials. Section 11.4 presents discussions and conclusion.

11.2 Predictive Accuracy and Clinical Trial Utility of Imaging Endpoints

11.2.1 Surrogate Endpoints for Trials

Approval of a new drug in USA requires substantial evidence of effectiveness that must be derived from adequate and well-controlled clinical investigations. Similarly, the Public Health Service Act requires biological products to be safe, pure, and potent. Clinical benefits that have supported drug approval have included important clinical outcomes (e.g., increased survival, symptomatic improvement) but have also included effects on established surrogate endpoints (e.g., blood pressure, serum cholesterol). Depending on the stage of drug development, clinical trials may use different endpoints to serve different purposes. For oncology drug development, for

example, early-phase clinical trials evaluate safety and identify evidence of biological drug activity, such as tumor shrinkage. Endpoints for later phase efficacy studies commonly evaluate whether a drug provides a clinical benefit such as prolongation of survival or an improvement in symptoms.

A clinical endpoint is defined as a characteristic or variable that reflects how patient feels or functions or how long a patient survives. They are strongest evidence for drug effects. Sometimes, it may take too long to measure a clinical endpoint. For that reason, we also use surrogate endpoint in trials. A surrogate endpoint is a biomarker intended to substitute for a clinical endpoint, which is expected to predict the clinical benefit, harm, or lack of benefit or harm. Imaging endpoints can be both clinical and surrogate endpoints.

Statistically, Prentice (1989) gave a statistical definition of surrogate endpoint "as a response variable (X) for which a test of the null hypothesis of no relationship to the treatment groups (T) under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint (Y)." Mathematically, it can be expressed as

$$f(X|T) = f(X) \Leftrightarrow f(Y|T) = f(T). \tag{11.1}$$

Here $f(Z|T)$ is the conditional probability distribution of a random variable Z conditional on the value of Z. Thus, $f(X|T) = f(X)$ represents the null hypothesis for the surrogate endpoint and $f(Y|T) = f(Y)$ represents the null hypothesis for the true clinical endpoint. "$\Leftrightarrow$" represents the equivalence.

Prentice criterion defined in (11.1) established the clinical validity of a surrogate marker for an intervention. Imaging endpoints measure clinically physiological changes invasively and can be seen at asymptomatic stage and can be perfect candidates for surrogate endpoints. However, this definition of surrogate endpoint requires equivalence, which is in a higher order than most commonly used surrogate endpoints (Schatzkin 2000). More often, treatment on surrogate endpoint is only a necessary but not sufficient condition for an effective treatment. For example, RECIST is a necessary condition for progress-free or overall survival for many oncology drugs. It is not a sufficient condition because some of the drugs may demonstrate benefits in RECIST but no improvement in overall survival. We often use RECIST as a surrogate endpoint for proof of concept phase II trials and progress survival time or overall survival time as clinical endpoint for the pivotal phase III trials. Also, it is worthwhile to point out that surrogacy of an endpoint depends on treatments.

Our discussion in this section is not about how to determine the clinical validity of an imaging endpoint as a surrogate endpoint (Alonso et al. 2006). We assume the case that the clinical validity has been established for imaging endpoints; i.e., the endpoint measures the clinical endpoints. To make a surrogate useful, in addition to its clinical validity defined by (11.1), we also need to establish its analytical validity and clinical utility. Analytical validity means that imaging biomarkers should be accurate, reliable, and reproducible, which we will discuss in Sect. 11.3. Clinical utility in a clinical trial setting means that use of surrogate endpoints should provide tangible benefits to investigators or sponsors, either in reduction of the number of

patients needed or trial durations or simplification of logistics, etc., to justify their uses. To this end, there is no difference between an imaging surrogate marker versus general surrogate markers. Therefore, the discussion below applies to all the surrogate markers.

11.2.2 Phase II Oncology Trials

In oncology trials, phase II trials often use a one-sided single-arm design. Let $Y(t)$ be a binary step function of the best clinical response status to proportion time t of trial duration since baseline: $t \in (0\,,\,1]$, such as complete or partial responses according to RECIST. We also assume that $Y(0) = 0$ as no response before treatment. In this section, we assume that a surrogate imaging endpoint X depends on treatment T through clinical outcome Y, i.e., $T \to Y \to X$. As the result, X is a surrogate endpoint for Y based on definition (11.1). Example of X can be circulated cancer DNA in blood sample. Because oncology phase II trials often are single arm, treatment T is not explicitly expressed in the experiment rather through the definitions of the null and alternative hypotheses. We want to determine if X has any clinical utility over endpoint Y.

Let $g(t) = Pr[Y(t) = 1]$ be the observed cumulative probability of response and $g^{(h)}(t) = Pr[Y(t) = 1|h\,]$ be the cumulative probability of response up to a proportion time of t under the null ($h = 0$) and alternative ($h = 1$) hypotheses, respectively. In a typical phase II trial design, we specified that a historical response rate as $g^{(0)}(1)$ and a success of a treatment should achieve a response rate at $g^{(1)}(1)$. Thus, a typical phase II trial test for the hypotheses that $H_0 : g(1) \le g^{(0)}(1)$ versus $H_1 : g(1) \ge g^{(1)}(1)$. The standard sample size formula for a single-arm one-sided phase II trial under is under a type I error α and power $1 - \beta$ is

$$n = \frac{\left(z_\alpha\sqrt{g^{(0)}(1)\big(1 - g^{(0)}(1)\big)} + z_{1-\beta}\sqrt{g^{(1)}(1)\big(1 - g^{(1)}(1)\big)}\right)^2}{\big(g^{(1)}(1) - g^{(0)}(1)\big)^2} \tag{11.2}$$

We can also consider the response rate at early time $Y(t)$ as a surrogate endpoint for $Y(1)$, the final response status. The specificity $P(Y(t) = 0|Y(1) = 0\,) = 1$ and sensitivity $P(Y(t) = 1|Y(1) = 1\,) = g(t)/g(1)$. In that sense, we can test if the drug is effective at any time t for $H_0 : g(t) \le g^{(0)}(t)$ versus $H_1 : g(t) \ge g^{(1)}(t)$. The asymptotic sample size is determined by

$$n(t) = \frac{\left(z_\alpha\sqrt{g^{(0)}(t)\big(1 - g^{(0)}(t)\big)} + z_{1-\beta}\sqrt{g^{(1)}(t)\big(1 - g^{(1)}(t)\big)}\right)^2}{\big(g^{(1)}(t) - g^{(0)}(t)\big)^2} \tag{11.3}$$

11.2.3 Continuous (Imaging) Surrogate Endpoints

Let X be a continuous (imaging) surrogate variable that can predict the response at the end of trial $Y(1)$. $X(t)$ is its value at time t. As a surrogate endpoint, the value of $X(t)$ for treatment responders and non-responders should have different distributions. Let

$$X_i(t) = \{X(t)|Y(1) = i\} \sim N\big(\mu_i(t), \sigma_i^2(t)\big) \tag{11.4}$$

be the surrogate variable for responders and non-responders. Without loss of generalizability, we let $\mu_1(t) - \mu_0(t) = \delta(t) > 0$. Thus,

$$X(t) = Y(1)X_1(t) + (1 - Y(1))X_0(t) \tag{11.5}$$

$$E[X(t)] = g(1)\mu_1(t) + (1 - g(1))\mu_0(t) = \mu_0(t) + g(1)\delta(t) \tag{11.6}$$

$$V[X(t)] = \sigma_0^2(t) + g(1)\big[\sigma_1^2(t) - \sigma_0^2(t)\big] + g(1)(1 - g(1))\delta^2(t) \tag{11.7}$$

The ability of $X(t)$ to predict response $Y(1)$ is often described by a receiver operating characteristics (ROC) curve. The area under the ROC curve (AUC) representing the strength of its predictive utility was studied in imaging validation studies. Under our notations, the AUC for $X(t)$ is measured by

$$AUC(t) = \Phi\left(\delta(t)/\sqrt{\sigma_0^2(t) + \sigma_1^2(t)}\right) \tag{11.8}$$

As a surrogate endpoint, test for the treatment efficacy at time t based on a test of mean surrogate measures at any time of t under the following hypotheses: H_0 : $E[X(t)] \leq \mu_0(t) + g^{(0)}(1)\delta(t)$ versus H_1 : $E[X(t)] \geq \mu_0(t) + g^{(1)}(1)\delta(t)$. The asymptotic sample size is determined by

$$m(t) = \frac{\left(\begin{array}{l} z_\alpha\sqrt{g^{(0)}(1)\big(1 - g^{(0)}(1)\big) + \frac{\sigma_0^2(t) + g^{(0)}(1)\left[\sigma_1^2(t) - \sigma_0^2(t)\right]}{\delta^2(t)}} + \\ z_{1-\beta}\sqrt{g^{(1)}(1)\big(1 - g^{(1)}(1)\big) + \frac{\sigma_0^2(t) + g^{(1)}(1)\left[\sigma_1^2(t) - \sigma_0^2(t)\right]}{\delta^2(t)}} \end{array}\right)^2}{\big(g^{(1)}(1) - g^{(0)}(1)\big)^2} \tag{11.9}$$

11.2.4 Discrete Imaging Surrogate Endpoints

Imaging surrogate endpoints can also be discretized. Let $Z(t)$ be a binary surrogate imaging endpoint. One possible example is to use a cutoff value of a continuous surrogate variable $X(t)$. We can define $Z(t|x) = 1_{\{X(t)>x\}}$. The sensitivity (Sn) and specificity (Sp) for clinical response $Y(1)$ are $Sn(t|x) = P(Z(t|x) = 1|Y(1) = 1) = 1 - \Phi\left(\frac{x-\mu_1(t)}{\sigma_1}\right)$ and $Sp(t|x) = P(Z(t|x) = 0|Y(1) = 0) = \Phi\left(\frac{x-\mu_0(t)}{\sigma_0}\right)$.

The mean and variance of $Z(t|x)$ are in (11.10) and (11.11), respectively.

$$h(t|x) = E[Z(t|x)] = [1 - Sp(t|x)] + [Sn(t|x) + Sp(t|x) - 1]g(1) \tag{11.10}$$

$$V[Z(t|x)] = g(1)(1 - g(1))[Sn(t|x) + Sp(t|x) - 1]^2 + Sp(t|x)(1 - Sp(t|x))(1 - g(1)) + Sn(t|x)(1 - Sn(t|x))g(1) \tag{11.11}$$

The sample size is calculated using the following formula.

$$k(t) = \frac{\left(z_\alpha \sqrt{\frac{g^{(0)}(1)\left(1 - g^{(0)}(1)\right) + \frac{Sp(t|x)(1-Sp(t|x))\left(1-g^{(0)}(1)\right) + Sn(t|x)(1-Sn(t|x))g^{(0)}(1)}{[Sn(t|x)+Sp(t|x)-1]^2}} + z_{1-\beta} \sqrt{\frac{g^{(1)}(1)\left(1 - g^{(1)}(1)\right) + \frac{Sp(t|x)(1-Sp(t|x))\left(1-g^{(1)}(1)\right) + Sn(t|x)(1-Sn(t|x))g^{(1)}(1)}{[Sn(t|x)+Sp(t|x)-1]^2}} \right)^2}{\left(g^{(1)}(1) - g^{(0)}(1)\right)^2}. \tag{11.12}$$

11.2.5 Utility of Surrogate Variables

For a surrogate variable to be useful, it should provide clinical utility for a trial. Here, we consider the utility in two aspects. The first utility is to reduce the overall sample size for a trial. The second utility is to allow a trial to reach conclusion earlier.

Theorem 11.1 *For a single-arm one-sided oncology trial with a binary clinical endpoint, either continuous or binary surrogate endpoints will not reduce the required sample sizes under the same type I or II errors, i.e.,* $m(t) \geq n(1)$ *and* $k(t) \geq n(1)$.

Proof According to Eqs. (11.2), (11.9), and (11.12), $m(t) \geq n(1)$ and $k(t) \geq n(1)$. Thus, surrogate endpoints will not be able to reduce the sample size during the entire trial period.

Furthermore, a continuous surrogate endpoint always requires higher sample size. For a binary surrogate variable, the sample size $k(t) = n(1)$ if and only if Z is a perfect surrogate with both sensitivity and specificity to be 1.

QED.

Definition 11.1 A surrogate variable has clinical utility value if it required a smaller sample size than $n(t)$ in Eq. (11.3) at a time t during the trial.

Thus, if $m(t)/n(t) < 1$ for a $t < 1$, the continuous surrogate endpoint has clinical utility value. Similarly, if $k(t)/n(t) < 1$ for a $t < 1$, the binary surrogate endpoint offers no clinical utility value.

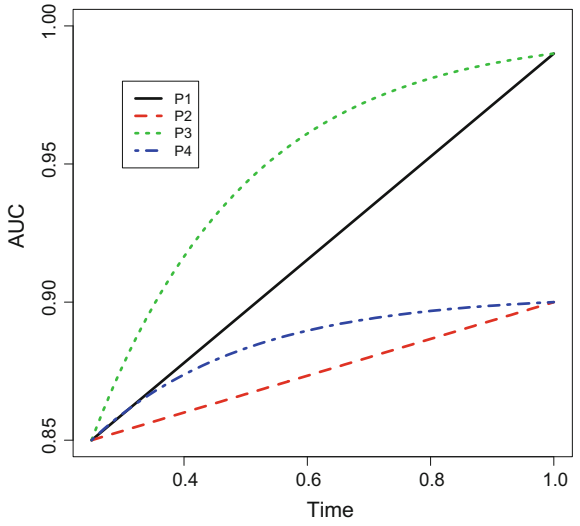


Fig. 11.1 AUC of ROC for P1 to P4 as a function of time during trial. Legend: The black (solid), red (dashed), green (dotted), and blue (dotdash) lines for P1, P2, P3, and P4, respectively

Whether a surrogate endpoint offers a clinical utility value depends on the relative improvement in its prediction of clinical response in comparison with the change of response over time. Using the ratio of $m(t)/n(t)$ and $k(t)/n(t)$ can help us to select surrogate endpoint that will reduce the trial duration. Furthermore, using $k(t)/n(t)$ can help us to select the best cutoff value for a continuous surrogate endpoint.

Example 11.1 Let $g^{(0)}(t) = 0.2t$ and $g^{(1)}(t) = 0.4t$, i.e., the response rates increase linearly under the null and alternative hypotheses for clinical endpoint. We have four continuous surrogate markers, $P1$, $P2$, $P3$, and $P4$. We assume that the accuracy of surrogate endpoint to predict clinical endpoint improves over the time. Using notations in Sect. 11.2.3, we assume that $\mu_0(t) = 0$ and $\sigma_0^2(t) = \sigma_1^2(t) = 1$. In this special case, $\mu_1(t) = \delta(t) = \sqrt{2}\Phi^{-1}(AUC(t))$. For P1 and P3, $AUC(1) = 0.99$ and for P2 and P4, $AUC(1) = 0.90$. After an initial period of no change, we assume AUCs of P1 and P3 follow a slow linear improvement with time as $AUC(t) = 0.85 + 4[AUC(1) - 0.85]/3(t - 0.25)_+$. P2 and P4, on the other hand, have exponential improvements in AUC as $AUC(t) = 0.85 + 1.05[AUC(1) - 0.85]\left[1 - exp(t - 0.25)_+\right]$. Figure 11.1 shows the AUCs for P1 to P4 as a function during the trial.

To evaluate type 2 clinical utility of these surrogate endpoints, Fig. 11.2 plots the ratio of sample size $m(t)/n(t)$ during trial period. Here, we use $\alpha = 0.05$ and $\beta = 0.2$. We can see that these markers will have advantage in early stage of trial but the benefits are reduced after more chance to observe clinical endpoints. The more accurate surrogate endpoints have benefits for a longer time window. Further, the rapidly improved endpoints have also longer clinical utility time windows.

Example 11.2 Continuing from Example 11.1, we would like to select a cutoff value in above continuous surrogate endpoints to generate binary surrogate endpoints. We

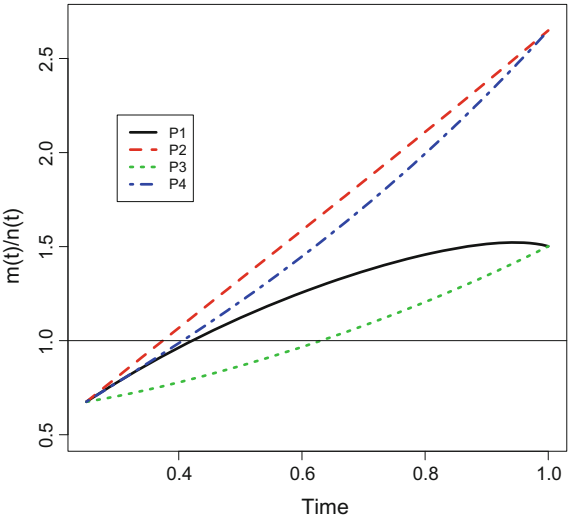


Fig. 11.2 Ratio of sample size $m(t)/n(t)$ during trial period. Legend: The black (solid), red (dashed), green (dotted), and blue (dotdash) lines for P1, P2, P3, and P4, respectively

Table 11.1 Comparison of clinical utility between continuous and optimal binary surrogates

Time	P1		P2		P3		P4	
	$\frac{m(t)}{n(t)}$	$\frac{k(t)}{n(t)}$	$\frac{m(t)}{n(t)}$	$\frac{k(t)}{n(t)}$	$\frac{m(t)}{n(t)}$	$\frac{k(t)}{n(t)}$	$\frac{m(t)}{n(t)}$	$\frac{k(t)}{n(t)}$
0.25	0.67	0.78	0.67	0.78	0.67	0.78	0.67	0.78
0.50	1.12	1.20	1.33	1.50	0.87	0.84	1.21	1.32
0.75	1.42	1.37	1.98	2.17	1.14	1.01	1.85	1.98
1.00	1.50	1.28	2.65	2.81	1.50	1.28	2.65	2.81

have many options to select the cutoff values. However, we would like to select the optimal cutoff values that generate lowest ratio of sample size $k(t)/n(t)$. The optimal cutoff values depend on the time during the trial. To illustrate selection of optimal cutoff values, we look at trial time $t = 0.25$, 0.50, 0.75, and 1.00 in Fig. 11.3. The left panel of Fig. 11.3 shows the ROC curves at four time points and highlights the optimal cutoff points. The right panel shows the ratio of $k(t)/n(t)$ over the range of 1-specificity and highlights the optimal cutoff points. Table 11.1 compares clinical utility values for all four surrogate measures used either as continuous or as binary at the optimal cutoff values. Although, in general, the continuous surrogate variables give better utility, it is possible that optimal binary surrogate endpoints outperform the continuous one in some cases.

11.2.6 Conclusion

To be used as surrogate endpoints for clinical trials, imaging endpoints need to have clinical relevance as defined by Prentice (1989) and clinical utility, either in

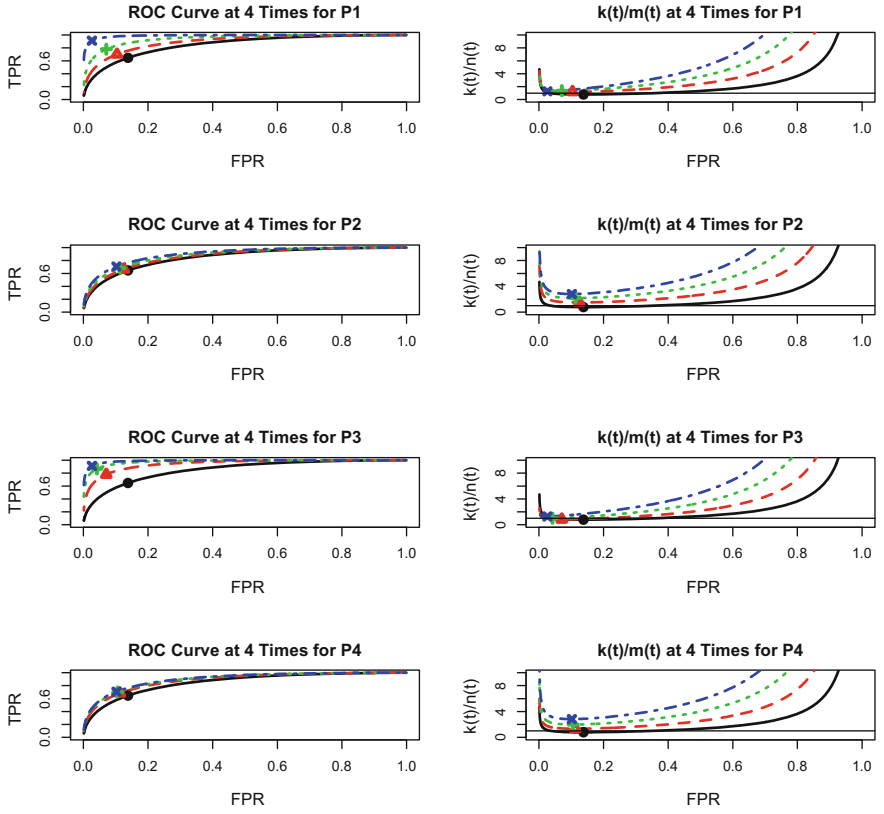


Fig. 11.3 Utility of binary surrogate endpoints and the corresponding optimal cutoff values at 0.25, 0.50, 0.75, and 1.00 time in a trial. Legend: The black (solid), red (dashed), green (dotted), and blue lines and points are for the 0.25, 0.50, 0.75, and 1.00 time point of a trial

saving overall sample size or saving time to conclude trial earlier. In this subsection, we demonstrated that surrogate endpoints will require more patients in a phase II single-arm oncology trial. It, however, can achieve earlier inference than the clinical endpoints if it has higher predictive accuracy measured in AUC of ROC curves and more rapidly increase in time than the clinical endpoints.

11.3 Quality Control Process for Imaging Endpoints

11.3.1 Types of Measurement Errors

Imaging endpoints depend on the condition of radiological equipment, operators, environment, and interpreters (radiologists). As such, their results are subject to random variations. Measurement error is defined as the difference between a measured value from the true physical value. Measurement errors can be characterized based on their statistical distribution. The difference between the mean of measurement errors and the true physical value is the *accuracy error*. The variation around the mean is the *precision error* (Lu and Zhao 2015).

Clinically, *accuracy errors* (here used as equivalent to the term bias) reflect the degree to which the measured results deviate from the true values. To evaluate accuracy errors, we need to know the true values of the measured parameters. It is not always possible, however, to measure the accuracy errors because sometimes the true values of the measured parameters cannot be verified. For example, quantitative ultrasound (QUS) bone measurements are affected by several quantitative and qualitative factors, and there is no single correlate for any QUS measurement. Therefore, we cannot define a single accuracy error for QUS (Njeh et al. 1999).

For clinical applications only the part of the accuracy error that varies from patient to patient in an unknown fashion is relevant. The other part, i.e., the one that is constant, can be averaged across subjects (e.g., the average underestimation of bone density due to the average fat content of bone marrow in quantitative computed tomography (QCT)), can be ignored. There are two reasons: First, for diagnostic uses, the reference data will be affected by the same error so the difference between healthy and diseased subjects is constant. Second, the error is present at both baseline and follow-up measurements and does not contribute to measured changes. Therefore, when discussing the impact of accuracy errors only that part of the error that changes from patient to patient in an unknown and uncontrollable fashion is of interest. For this reason, small accuracy errors are of little clinical significance provided they remain constant. They are more relevant to diagnosis and risk assessment than to monitoring changes or relative differences between arms in a clinical trial (Engelke and Gluer 2006).

Precision errors reflect the reproducibility of the technique. They measure the ability of a method to reproducibly measure a parameter for reliably monitoring clinical changes over time. Precision errors can be further separated into *short-term* and *long-term* precision errors. Short-term precision errors characterize the reproducibility of a technique and are useful for describing the limitations of measuring changes in a clinical status. If they are large, they may affect the diagnostic accuracy of a technique measured by AUC of ROC curves. Long-term precision errors are used to evaluate instrument stability which is critical for clinical trials. Because long-term precision errors include additional sources of random variation attributable to small drifts in instrumental calibration, variations in patient characteristics, and other technical changes related to time, they provide a better measure of a technique's ability

to monitor parameter changes than the short-term precision errors do. For patient measurements, estimates of long-term precision usually also include true longitudinal variability of measured body sites. For these reasons, long-term precision errors normally are larger than short-term errors. Lu and Zhao (2015) provided a comprehensive overview on the statistics used in quality control, quality assurance, and quality improvement in radiological studies, including various forms of precision errors and their distribution properties for different clinical applications.

Precision errors are usually evaluated whenever new techniques or new devices are developed. Precision errors are also evaluated immediately after a device is installed in clinical sites to assure that the equipment is performing according to the manufacturer's specifications at baseline. Precision errors also are always assessed before the beginning of clinical trials or longitudinal studies (Lu et al. 1996). It is always the desire to keep the same equipment used through the entire trial to assure consistency of measurements. However, for long-term studies, it is usually not feasible. Although the manufacturer's service personnel can set up the device so that precision errors are within appropriate limits at baseline, it is very important to monitor the equipment to assure that imprecision remains within acceptable limits. Despite the remarkable accuracy and reproducibility of radiological equipment, measurements can still vary because of changes in equipment, software upgrades, machine recalibration, X-ray source decay, hardware aging and/or failure, or operator errors. If equipment is upgraded during the trial, in addition to calibrate to the previously used device, a new precision reference is usually established. A conversion formula (Shepherd and Lu 2007) can be used to characterize the precision errors of longitudinal changes in the presence of device upgrades.

A quality control and quality assurance program is always necessary for clinical trials using imaging endpoints. In this section, we would like to provide overview of extension of univariate statistical process control approaches (Lu and Zhao 2015) to bivariate quality control process charts, which have been used in pediatric trials to monitor bone mineral density and contents over the trial period of time.

11.3.2 Univariate Process Control Charts for Bone Mineral Density

In an ideal setting, a well-maintained equipment produces values that are randomly spread around a reference value. A change point is defined as the point in time at which the measured values start to deviate from the reference value. To evaluate measurement stability and identify change points, radiologists develop phantoms that simulate human measurements but, unlike humans, do not change over time (Kalender et al. 1995; Krueger et al. 2016). Variations in phantom measurements should reflect variations in human measurements. Phantoms are measured regularly to detect one or more of the following events: (1) The mean values before and after the change point are statistically significantly different; (2) The standard deviations

of measurements before and after the change point are statistically significantly different; (3) The measurements after the change point show a gradual but significant departure from the reference value. When phantom is not available and possible, normal health controls have been traveled to all participating sites to serve as the references longitudinally and across sites (Keshavan et al. 2016). There are advantages of in vivo measures of human subjects. However, it costs a lot more, cannot continuously measured, and is often hard to determine if there are changes over time for the reference populations. Therefore, if possible, a quality control based on phantoms is the best approach for clinical trials.

Statistical process control (SPC) is a powerful collection of problem-solving tools for achieving process stability and improving capacity through reduction of variability (Rodriguez and Ransdell 2010). There are several statistical methods for identifying change points (Montgomery 2012). Shewhart chart and CUSUM chart are two most commonly used methods. The advantages of Shewhart chart are its simplicity and easy for use. Thus, it has been widely used in many areas, especially direct use by technicians of radiology equipment. CUSUM is a more sensitive statistical method that can identify status of an equipment more accurately and efficiently. However, it requires mathematical calculations and much less intuitive. It is often used by quality control centers to monitor machine performance during a trial.

In osteoporosis trials using DXA scanners, Shewhart and CUSUM process control charts are used to monitor longitudinal scanner performance for BMD either using Hologic or European spine phantoms or Hologic hip or whole-body phantoms. Details of these quality control charts and examples of their applications for univariate BMD can be found in Lu et al. (1996), Lu and Zhao (2015). Here, we give a brief overview in order to introduce the bivariate quality control charts.

A Shewhart chart is a graphic display of a quality that has been measured over time. The chart contains a central horizontal line that represents the mean reference value. Three horizontal lines above and three below the central line indicate 1, 2, and 3 standard deviations from the reference value. By plotting the observed quality control measurements on the chart, we can determine if the machine is operating within acceptable limits.

The reference values can be derived from theoretical values for the phantom, or from the first 25 observations measured at baseline. The reference value changes whenever the Shewhart chart indicates an out-of-control signal and the machine is recalibrated. The new reference value will then be the mean of the first 25 observations after recalibration. The number of observations needed to calculate the reference value may vary; the number 25 was chosen based on practical experience to balance the stability of the reference value with the length of time needed to establish it.

The standard deviation varies among individual devices, and manufacturers should be selected accordingly. For example, in one osteoporosis study, we sometimes use the BMD of a Hologic phantom to monitor DXA scanner performance. We usually assume the coefficient of variation for Hologic machines to be 0.5% and Lunar to be 0.6%, based on reported data on long-term phantom precision (Lu et al. 1996). Therefore, the standard deviation for the scanner was calculated as 0.005 and 0.006 times the reference value for Hologic and Lunar machines, respectively.

Table 11.2 Definition of tests for assignable causes for Shewhart charts

Tests	Pattern description
1	One point is more than 3 standard deviation from the central line
2	Nine points in a row on one side of the central line
3	Six points in a row steadily increasing or steadily decreasing
4	Fourteen points in a row alternating up and down
5	Two out of 3 points in a row more than 2 standard deviation from the central line
6	Four out of 5 points in a row more than 1 standard deviation from the central line
7	Fifteen points in a row all within 1 standard deviation from the central line on either or both sides of the line
8	Eight points in a row all beyond 1 standard deviation from the central line on either or both sides of the line

The original Shewhart chart will signal that there is a problem if the observed measurement is more than 3 standard deviations from the reference value. Although intuitive and easy to apply, the chart is not very sensitive to small but significant changes. Therefore, a set of sensitizing tests for assignable causes has been developed to improve the sensitivity of Shewhart charts. Eight of the tests are available in the statistical software package SAS. The tests are listed in Table 11.2.

The sensitizing rules can be used in all or in part depending on the underlying processes of interest. For example, for quality control of DXA machines, we used four tests—1, 2, 5, and 6 (Lu et al. 1996). Once a change point has been identified by any one of the tests, the manufacturer's repair service should be called to examine the causes and to recalibrate the machine. We then use the next 25 observations to generate new reference values and apply the tests to the subsequent data according to the new reference value.

The sensitizing rules increase the sensitivity of the Shewhart chart, but also increase the number of clinically insignificant alarms, which is not desirable. To overcome this problem, a threshold based on the magnitude of the mean shift can also be implemented. For example, we can select ten consecutive scans from after the possible change point identified on the Shewhart chart and then calculate their mean values. If the mean differs by more than one standard deviation (which equals 0.5% times the reference value, in our example) from the reference value, the change point is confirmed as a true change point. Otherwise, the signal from the Shewhart chart is ignored and the reference value is unchanged. This approach filters out small and clinically insignificant changes. However, the true difference must be more than one standard deviation for this approach to be effective, and this approach can delay the recognition of true change points.

A CUSUM chart is short for cumulative sum chart. In applications, we recommend a version of CUSUM known as tabular CUSUM because it can be presented with or without graphs. Mathematically, we define an upper one-sided tabular CUSUM $S_H(i)$ and a lower one-sided tabular CUSUM $S_L(i)$ for the ith QC measurement as the following:

$$S_H(i) = \max\left[0, \frac{X_i - \mu_0}{\sigma} - k + S_H(i-1)\right] \tag{11.13}$$

$$S_L(i) = \max\left[0, \frac{\mu_0 - X_i}{\sigma} - k + S_L(i-1)\right] \tag{11.14}$$

Here, μ_0 is the reference mean, σ is the standard deviation, and k is a parameter to filter out insignificant variations and is usually set at 0.5. The initial values of $S_H(0)$ and $S_L(0)$ are 0. The chart sends an alarm message if $S_H(i)$ or $S_L(i)$ is greater than 5. In other words, when the standardized BMD value deviates more than k from zero, the cumulative upper bounded sum increases by the amount of deviations above k. On the other hand, if the deviation is less than k, the cumulative sum will be reduced accordingly. When the cumulative sum is less than zero, we ignore the past data and set the cumulative sum as zero. However, a cumulative sum greater than 5 is a strong indication of a deviation from the reference mean in the data. Furthermore, a CUSUM chart also estimates when the change occurred and the magnitude of the change. We use the estimated magnitude of change to establish the new reference values (Lu and Zhao 2015).

11.3.3 Bivariate Process Control Charts for Bone Mineral Density

Bone structure for adults does not change rapidly, and bone mineral density change is primarily due to the loss of bone mineral contents (BMC). Thus, using longitudinal Shewhart or CUSUM chart described above is adequate to monitoring DXA scanner performance in adult osteoporosis trials. The situation is different, however, for pediatric trials, during which children experience rapid changes not only in bone mineral contents but also in bone structure due to growth. Thus, quality control is needed to monitor the measurement stability in not only BMD, but also both BMC and area for whole-body scans in pediatric clinical trials for bone changes. As such, a bivariate simultaneous process control chart is required for BMB and BMC. Because BMD is the ratio of BMC over bone area, in control of BMD and BMC implies in control of the bone area. In our bivariate quality control process, we perform log-transformations for BMD (*l*BMD) and BMC (*l*BMC), so the log-transformed bone area is the difference of the *l*BMC and *l*BMD. In ten sets of quality control scan data for whole-body phantom in a pediatric trial we tested, the log-transformed BMD, BMC, and Area had almost identical break points as the raw data of BMD, BMC, and Area. Thus, in this section, log-transformed variable names are used exchangeable with the original BMD, BMC, and Area. We introduce bivariate Shewhart and CUSUM charts as the following.

Bivariate Shewhart Chart (BSC) (Lu et al. 2006): In a univariate Shewhart chart, we draw three lines above and below the reference mean in a unit of one standard deviation. Based on the probability of a normally distributed random variable, we

Table 11.3 Bivariate Shewhart sensitized rules for QC Data

Rules	Description
1	One (1) point outside the outer ring (third ring)
2	Five (5) points in a row on one of the four quatrains
5	Two (2) of 3 points in a row outside of the middle ring (second ring)
6	Four (4) of 5 points in a row outside of the inner ring (first ring)

define several types of reportable events in Table 11.2. In our bivariate Shewhart chart, we assume the random measurement errors for *l*BMD and *l*BMC follow a bivariate normal distribution with a mean of (μ_{lBMD}, μ_{lBMC}) and a covariance matrix Σ. We denote the standard deviations of *l*BMD and *l*BMC as σ_{lBMD} and σ_{lBMC}, and the correlation coefficient as ρ. Like univariate case, we can estimate these five parameters from the first 25 QC phantom scan data.

The main idea to extend Shewhart chart form univariate to bivariate is to divide the bivariate plane into different rings and define events accordingly. Let *l*BMC be the X-axis and *l*BMD be the Y-axis. We draw three ellipses using the following equation:

$$\sigma_{lBMD}^2(X-\mu_{lBMC})^2 - 2\rho\sigma_{lBMD}\sigma_{lBMC}(X-\mu_{lBMC})(Y-\mu_{lBMD}) + \sigma_{lBMC}^2(Y-\mu_{lBMD})^2 = \sigma_{lBMD}^2\sigma_{lBMC}^2\left(1-\rho^2\right)R^2 \quad (11.15)$$

where R = 1.52, 2.49, and 3.44, respectively, to reflect similar boundaries of one, two, and three standard deviation lines in a univariate Shewhart chart.

We then draw the bivariable plane into four quatrains according to two lines by the axes of the ellipses. Figure 11.4 shows steps to construct the four quatrains. A 45-degree line ($lBMC = lBMD$, or $y = x$) can be also drawn to indicate the directions of measured area in relationship to the observed reference means. If observed break points moved toward northeast, it means the area increased.

Whenever the machine recalibrated or maintained, a new chart should be formed based on the new means and covariance matrix.

Table 11.3 gives sensitizing rules for bivariate Shewhart chart that are corresponding to Table 11.2 for univariate Shewhart chart rule. After daily quality control scan, a technician plots the *l*BMD and *l*BMC on the bivariate Shewhart chart. If one of the above four rules is violated, the machine will need to be reviewed for assignable courses and possibly required for services.

As an example of a whole-body phantom quality scans by a DXA scanner, we calculated five parameters for *l*BMD and *l*BMC: $\mu_{lBMD} = 0.1008$ (log(mg/cm^2)), $\mu_{lBMC} = 6.5814$ (log(mg)), $\sigma_{lBMD} = 0.0256$ (log(mg/cm^2)), $\sigma_{lBMC} = 0.0320$ (log(mg)), and $\rho = 0.5885$. Figure 11.5 shows the corresponding Shewhart chart and examples when the rules are violated.

Bivariate CUSUM Chart (BCC) (Lu et al. 2007): To construct bivariate CUSUM, we first transform the *i*th observed *l*BMC (X_i) and *l*BMD (Y_i) into two independent uncorrelated standardized monitoring variables $Z_i^{(1)}$ and $Z_i^{(2)}$. Here,

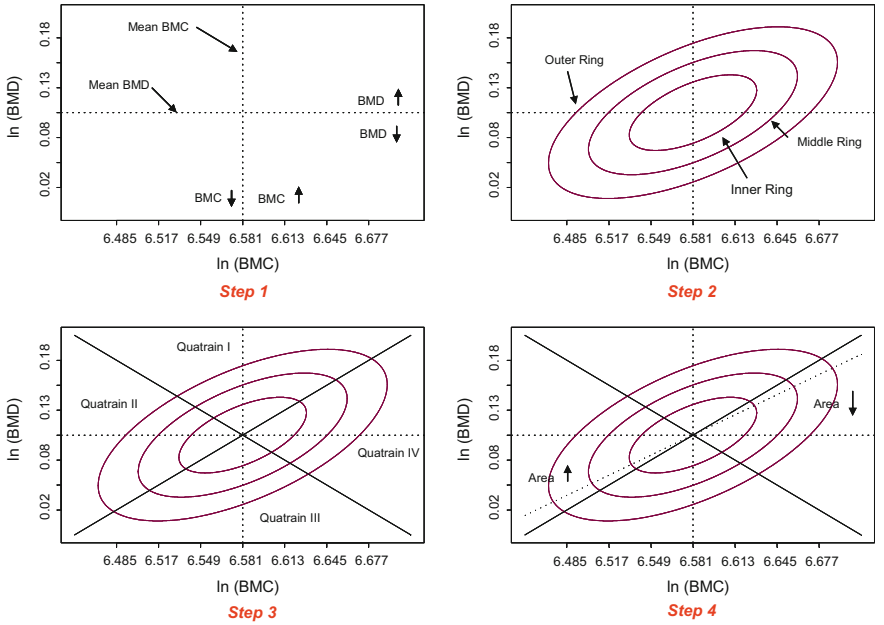


Fig. 11.4 Illustration of steps to construct a bivariate Shewhart chart. Legend: • Step 1: Take first 25 QC phantom data, and calculate the 5 baseline distributions. • Step 2: Create a chart using lBMC as the X-axis and lBMD as the Y-axis. Draw three ellipses using the following equations: $\sigma_{lBMD}^2(X-\mu_{lBMC})^2 - 2\rho\sigma_{lBMD}\sigma_{lBMC}(X-\mu_{lBMC})\,(Y-\mu_{lBMD}) + \sigma_{lBMC}^2(Y-\mu_{lBMD})^2 = \sigma_{lBMD}^2\sigma_{lBMC}^2\left(1-\rho^2\right)R^2$. • Three *R*s to be used are 1.52, 2.49, and 3.44, respectively, to reflect similar boundaries of three SD lines in a univariate Shewhart chart.• Step 3: A 45-degree line ($Y=X$) can be drawn to indicate the directions of measured area in relationship to the observed reference mean.• Step 4: Divide the plane into four quatrains according to two lines by the axes of the ellipses.• A chart organized in Steps 1–4 will be kept over time until a break point is identified and the scanner recalibrated. A new chart should start then

$$\begin{bmatrix} Z_i^{(1)} \\ Z_i^{(2)} \end{bmatrix} = \begin{bmatrix} \sigma_{lBMC}^2 & \rho\sigma_{LBMC}\sigma_{LBMD} \\ \rho\sigma_{LBMC}\sigma_{LBMD} & \sigma_{lBMD}^2 \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} X_i - \mu_{ilBMC} \\ Y_i - \mu_{ilBMD} \end{bmatrix} \tag{11.16}$$

Then, we can construct the following cumulative sums for the deviation from the origin (0, 0):

Step 1: Take first 25 QC phantom data, and calculate the means (μ_{lBMC}, μ_{lBMD}), standard deviations (σ_{lBMC}, σ_{lBMD}) and correlation coefficient ρ between *l*BMC and *l*BMC.

Step 2: Transform the *i*th observed *l*BMC and *l*BMD in longitudinal QC data into two uncorrelated standardized monitoring variables $Z_i^{(1)}$ and $Z_i^{(2)}$ according to Eq. (11.16).

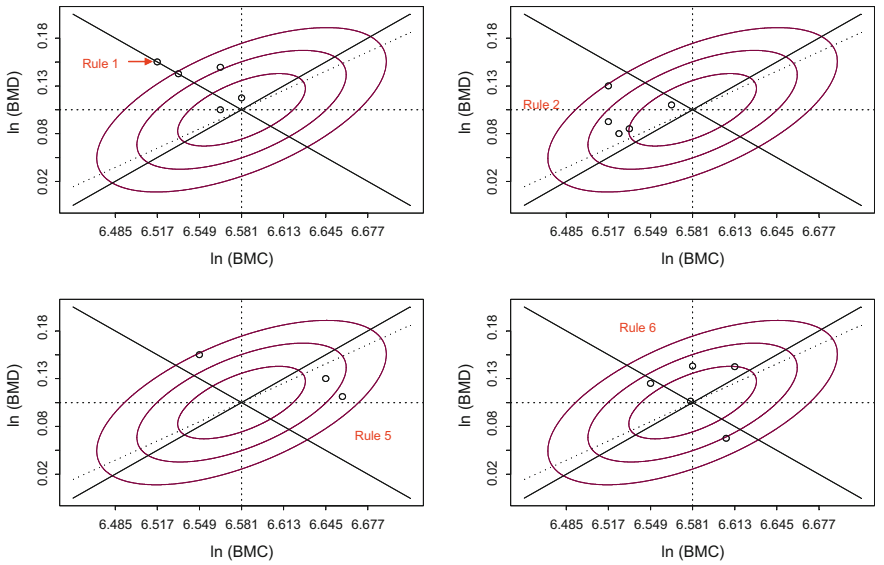


Fig. 11.5 Example of alarm events in a bivariate Shewhart control chart for BMD data

Step 3: Calculate distance from center (0,0) as $R_i = \sqrt{\left(Z_i^{(1)}\right)^2 + \left(Z_i^{(2)}\right)^2}$

Step 4: Calculate the cumulative upper and lower sums of the ith observation in both in $Z_i^{(1)}$ and $Z_i^{(2)}$ using the following formulas:

$$S_H^{(1)}(i) = \max\left(0, S_H^{(1)}(i-1) + (R_i - K)\frac{Z_i^{(1)}}{R_i} 1_{\left\{Z_i^{(1)} > 0\right\}}\right) \tag{11.17}$$

$$S_L^{(1)}(i) = \max\left(0, S_L^{(1)}(i-1) + (K - R_i)\frac{Z_i^{(1)}}{R_i} 1_{\left\{Z_i^{(1)} < 0\right\}}\right) \tag{11.18}$$

$$S_H^{(2)}(i) = \max\left(0, S_H^{(2)}(i-1) + (R_i - K)\frac{Z_i^{(2)}}{R_i} 1_{\left\{Z_i^{(2)} > 0\right\}}\right) \tag{11.19}$$

$$S_L^{(2)}(i) = \max\left(0, S_L^{(2)}(i-1) + (K - R_i)\frac{Z_i^{(2)}}{R_i} 1_{\left\{Z_i^{(2)} < 0\right\}}\right) \tag{11.20}$$

Here, subscripts H and L indicate the higher and lower directions of cumulative deviations from the means; (i) indicates variable number (1 and 2); K is the filter for deviation and has been selected as 0.5; and finally, $1_{\{G\}}$ is the indication function of event G being true (=1) or false (=0).

Step 5: Send an alarm if one of the sum is above B.

11.3.4 Comparison of Bivariate Quality Control Process Charts and Combinational Uses of Univariate Quality Control Charts

In this subsection, we compare the performance of univariate QC charts, simultaneous uses of univariate quality control charts, and bivariate QC charts for both BMD and BMC. The first approach simultaneously uses Shewhart charts for *l*BMD and *l*BMC to monitor scanner performance. If one or both two variables violate Shewhart rules 1, 2, 5, and 6 in Table 11.2, the algorithm will send an alarm. The second approach simultaneously uses CUSUM charts for *l*BMD and *l*BMC. An alarm is set if one or both CUSUM detect an alarm. The third approach uses bivariate Shewhart chart, and the fourth approach uses bivariate CUSUM. We refer them as Approaches 1–4, respectively, in this section.

Simulation conditions are set in Table 11.4. The simulation data are generated in two period. Period 1 is for case when the scanner is in control and is used to evaluate performance in control conditions. The scanner changes in Period 2. In the first scenario, both BMD and BMD increase 1SD (and thus *l*Area is unchanged). In the second scenario, both BMD and BMD increase 0.707SD (and thus *l*Area is unchanged). In the third scenario, only BMC increases 1SD (and thus *l*Area also increases the same amount). In the fourth scenario, BMC increases 1SD and BMD decreases 1SD (and thus *l*Area has a 2SD decrease). We measure the performance based on two variables. The first is the average running length (ARL) from the start to the first alarm. When the system is in control, we want the average running length as long as possible. However, when the system changes, we want the average running length as short as possible. The second outcome variable is the rate of sending an alarm within 100 consecutive scans. When the process is in control, this rate is corresponding to the false positive rate within 100 scans (FPR), which should be as low as possible. When the process is not in control, this rate is corresponding to a true positive rate (TPR) and should be as high as possible. Table 11.5 gives the simulation results.

From Table 11.5, we can see that bivariate CUSUM has the best performance with excellent sensitivity while having high specificity. Twenty-five percent (25%) of simulations had no false positives in 365 scans. In comparison, the bivariate Shewhart chart has also good sensitivity, not high false alarm rate. Similarly, the combinational use of univariate CUSUM achieved high sensitivity but not specific enough as the false alarm rates were high in all four cases. Combination of two Shewhart was not sensitive enough in scenario 2, and the specificity is also worse than CUMSUM but better than bivariate Shewhart chart.

We applied these methods to longitudinal QC data from four DXA scanners from four different sites. Three scanners monitored the whole-body (WB) phantoms, and one scanner monitored a spine phantom (SP). Table 11.6 lists the number of scans to reach the first alarm using Shewhart or CUSUM chart to monitor BMD and BMC. In addition to the univariate QC, we also compare the alarm by either BMD or BMC whatever first, or joint bivariate QC charts. Although we did not have gold

Table 11.4 Distribution of *l*BMC and *l*BMD in four simulation scenarios

Scenarios	Period 1 (n = 365)	Period 2 (n = 365)
1	$N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5885\\0.5885 & 1\end{bmatrix}\right)$	$N\left(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1 & 0.5885\\0.5885 & 1\end{bmatrix}\right)$
2	$N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5885\\0.5885 & 1\end{bmatrix}\right)$	$N\left(\begin{bmatrix}0.707\\0.707\end{bmatrix}, \begin{bmatrix}1 & 0.5885\\0.5885 & 1\end{bmatrix}\right)$
3	$N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5885\\0.5885 & 1\end{bmatrix}\right)$	$N\left(\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5885\\0.5885 & 1\end{bmatrix}\right)$
4	$N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5885\\0.5885 & 1\end{bmatrix}\right)$	$N\left(\begin{bmatrix}1\\-1\end{bmatrix}, \begin{bmatrix}1 & 0.5885\\0.5885 & 1\end{bmatrix}\right)$

Table 11.5 Average running length and true and false positive rates of simulation experiments

Condition	Statistics	Univariate CUSUM			Simultaneous approaches			
		BMD	BMC	Area	1	2	3	4
In control	ARL[a]	156	157	158	80	61	31	213
	FPR[b]	37%	38%	38%	57%	69%	91%	29%
Scenario 1	ARL[a]	18	17	80	6	5	6	8
	TPR[c]	74%	74%	55%	91%	95%	98%	99%
Scenario 2	ARL_a	115	19	18	11	5	6	7
	TPR[c]	47%	73%	77%	85%	95%	98%	100%
Scenario 3	ARL[a]	21	19	96	8	7	9	20
	TPR[c]	79%	79%	51%	95%	97%	98%	88%
Scenario 4	ARL[a]	17	20	24	6	4	5	5
	TPR[c]	81%	80%	71%	96%	98%	98%	100%

[a]ARL (average running length) is the number of scans until the first alarm
[b]FPR (false positive rate) is the percentage of simulations that the QC method gave a false positive alarm within 100 scans after the beginning of the monitoring
[c]TPR (true positive rate) is defined as the percentage of simulations that the QC method gives a true positive alarm within 100 scans after the event

standards (i.e., the maintenance records for assignable courses) for these scanners, the first scanner may have some issues with BMC. It is almost for sure that the second scanner had a failure in all parameters. The scanner 3 most likely is in well-controlled condition, and the fourth scanner may have some problems with BMD. Except for scanners 1 and 3, which may be in control condition, the performance using Shewhart or CUSUM chart has no practical differences.

Table 11.6 Applications of bivariate QC to four QC data

Site/Phantom	Methods	Univariate QC			Simultaneous approaches	
		BMD	BMC	Area	BMD or BMC	Bivariate QC
1/WB	Shewhart	17	19	31	17	15
	CUSUM	26	16	131	26	76
2/WB	Shewhart	5	5	14	5	5
	CUSUM	2	6	2	2	2
3/WB	Shewhart	57	70	9	57	29
	CUSUM	63	258	258	63	69
4/SP	Shewhart	16	14	48	14	16
	CUSUM	20	45	96	20	54

11.3.5 Conclusion

For the validity of clinical trial results, quality control programs are often required for clinical trials using imaging endpoints. Statistical quality control methods, such as Shewhart and CUSUM charts, are used in trials. Increasingly, QC requires monitoring multiple imaging measures. Repeated use of univariate QC process control charts increases chance of false alarms. A proposed bivariate CUSUM chart demonstrated outstanding performance characteristics in DXA QC for pediatric trials that required simultaneous monitoring of both BMC and BMD.

11.4 Discussions and Conclusions

Imaging biomarkers have been widely used in clinical trials, both as surrogate endpoints or diagnostic tools to select proper patient populations. The use of imaging endpoints is continuously increasing. The use of new imaging technology in clinical trials has raised many statistical questions. This chapter focused on a small part of evaluating the utility of surrogate markers for clinical trials and the quality control methods to monitor performance of radiological equipment during clinical trials.

We demonstrated that the usefulness of a surrogate endpoint relative to direct use of clinical endpoint depends on the leading time of the biomarkers relative to the clinical endpoints and diagnostic precision. A binormal model is a nature way to link the change over time of the treatment effect size of the surrogate marker to a commonly used diagnostic efficiency parameter, AUC of a ROC curve. Modeling the time to event distribution and the leading time AUC will help us to assess whether using a surrogate marker offers advantages in a trial. We are extending this approach to randomized clinical trials with different types of clinical endpoints.

We also illustrated the use of process control charts to monitor bivariate measures in pediatric bone mineral density studies. Another important question for quality control is the calibration among different equipments from different sites. Lu and Zhao (2015) has extensive discussions on this topic. Keshavan and colleagues (2016) evaluated the impact of cross-calibration on statistical power in multisite MRI studies for multiple sclerosis trials.

Acknowledgements I would like to make acknowledgment of contributions of my colleagues and funding supports. The bivariate QC was a joint work by Drs. John Shepherd, Shoujun Zhao, and Bo Fan at the Department of Radiology and Biomedical Imaging, University of California, San Francisco. That part of work was supported by a grant from CDC 200-2005-11219 (PI: Dr. Shepherd). The first part of utility of surrogate endpoints was supported by a grant from NIH R01 EB004079-01 (PI: Lu). The work was performed when author worked at the University of California, San Francisco.

References

Alonso, A., Molenberghs, G., Geys, H., Buyse, M., & Vangeneugden, T. (2006). A unifying approach for surrogate marker validation based on Prentice's criteria. *Statistics in Medicine, 25*(2), 205–221.

Engelke, K., & Glüer, C. C. (2006). Quality and performance measures in bone densitometry. Part 1: Errors and diagnosis. *Osteoporosis International, 17*(9), 1283–1292.

Hillman, B. J. (2005). ACRIN—Lessions learned in conducting multi-center trials of imaging and cancer. *Cancer Imaging, 5*, Spec No A:S97-101. PMID: 16361142.

Kalender, W. A., Felsenberg, D., Genant, H. K., Fischer, M., Dequeker, J., & Reeve, J. (1995). The European Spine Phantom—A tool for standardization and quality control in spinal bone mineral measurements by DXA and QCT. *European Journal of Radiology, 20*(2), 83–92.

Keshavan, A., Paul, F., Beyer, M. K., et al. (2016). Power estimestimationatino for non-standardized multisite studies. *Neuroimage, 134,* 281–294. https://doi.org/10.1016/j.neuroimage.2016.03.051. Epub 2016 Apr 1.

Krueger, D., Libber, J., Sanfilippo, J., Yu, H.J., Horvath, B., Miller, C. G., et al. (2016) A DXA whole body composition cross-calibration experience: Evaluation with humans, spine, and whole body phantoms. *Journal of Clinical Densitometry, 19*(2), 220–225.

Lu, Y., Mathur, A. K., Blunt, B. A., Gluer, C. C., Will, A. S., Fuerst, T. P., et al. (1996). Dual X-ray absorptiometry quality control: Comparison of visual examination and process-control charts. *Journal of Bone and Mineral Research, 11*(5), 626–637.

Lu, Y., & Zhao, S. (2015). Statistics used in quality control, quality assurance, and quality improvement in radiological studies. In Y. Lu, J. Fang, L. Tian, & H. Jin (Eds.), *Advanced medical statistics* (pp. 103–160). New York: World Scientific.

Lu, Y., Zhao, S., Fan, B., & Shepherd, J. (2006). Simultaneous process control charts for BMD, BMC, and Area in longitudinal quality control of DXA scanners. In *15th International Bone Densitometry Workshop*, Kyoto, Japan, Oct 2006.

Lu, Y., Zhao, S., Fan, B., & Shepherd, J. (2007) A new CUSUM method for simultaneous quality control of BMD, BMC, and Area for DXA scanners. In *29th Annual Meeting of American Society of Bone and Mineral Research*, Honolulu, Hawaii, USA, 2007.

Mongomery, D. C. (2012). *Introduction to statistical quality control* (7th ed.). New York: Wiley.

Njeh, C. F., Richards, A., Boivin, C. M., Hans, D., Fuerst, T., & Genant, H. (1999). Factors influencing the speed of sound through the proximal phalanges. *Journal of Clinical Densitometry, 2*(3), 241–249.

No authors listed. (1948). STREPTOMYCIN treatment of pulmonary tuberculosis. *British Medical Journal, 2,* 24.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine, 8,* 431–440.

Rodriguez, R. N., & Ransdell, B. (2010). *Statistical process control for heath care quality improvement using SAS/QRobert N*. Cary, NC: SAS Institute Inc.

Schatzkin, A. (2000). Intermediate markers as surrogate endpoints in cancer research. *Hematology/Oncology Clinics of North America, 14*(4), 887–905.

Shepherd, J. A., & Lu, Y. (2007). A generalized least significant change for individuals measured on different DXA systems. *Journal of Clinical Densitometry, 10*(3), 249–258.

Thomas, A. M. K., & Banerjee, A. K. (2013). *History of radiology: Oxford medical history*. Oxford: Oxford University Press.

Zhao, Q., et al. (2010). A statistical method (cross-validation) for bone loss region detection after spaceflight. *Australasian Physical and Engineering Sciences in Medicine, 33*(2), 163–169. https://doi.org/10.1007/s13246-010-0024-6. Epub 2010 Jul 15.

Chapter 12
Interesting Applications from Three Decades of Biostatistical Consulting

Karl E. Peace, Uche Eseoghene Okoro and Kao-Tai Tsai

12.1 Introduction

In 30 years of consulting in and to the pharmaceutical industry, several novel and challenging applications arose. Four such examples are presented in this chapter. The example presented in Sect. 12.2 demonstrates using a parallel design instead of a crossover design in a bioequivalence trial of several formulations. This trial also incorporated blinded sample size re-estimation—before there was a literature on the subject.

The second example, presented in Sect. 12.3, is of a clinical trial employing a 2×2 factorial design in studying the effects of a fixed combination, a two-component drug for treating allergic rhinitis. This trial reflects using bivariate plots of two primary response measures to illustrate simultaneously drug/dose effects.

The third example, presented in Sect. 12.4, is a dose comparison trial aimed at establishing the optimal dose of a H2-receptor antagonist in treating patients with duodenal ulcer. This trial also reflects using bivariate plots of two primary response measures to illustrate simultaneously drug/dose effects, and reflects logistical aspects of interim analyses procedures aimed initially at dropping the placebo arm and later dose discrimination, as well as illustrating the need to carefully define the trial objective.

The fourth and last example, presented in Sect. 12.5, provides an overview of assessing whether evidence exists from two Phase II trials of angina to support conducting a Phase III trial at either a b.i.d. or t.i.d. dosing of the drug. This example reflects using an equiradial hexagonal design and response surface methods to arrive at a region in the dose by frequency of dosing plane over which maximal response is expected.

K. E. Peace (✉) · U. E. Okoro · K.-T. Tsai
Jiann-Ping Hsu School of Public Health, Georgia Southern University, 30461 Statesboro, Georgia
e-mail: peacekarl@frontier.com; kepeace@georgiasouthern.edu

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_12

12.2 A Parallel Design Instead of a Crossover Design in a Bioequivalence Study of Several Formulations (Peace and Chen 2010)

12.2.1 Background

A 2×2 crossover bioequivalence trial of a new test formulation (T1) versus the standard (S) of a marketed drug had been conducted in 16 subjects. One subject failed to complete the second period. Although a washout period of at least three half-lives of the drug was included in the protocol, analysis of AUC and CMAX suggested the presence of a carry-over effect (or treatment-by-period interaction). Therefore, inference as to bioequivalence had to rely on only the data from the first period. The analysis of first period bioavailability endpoints did not enable one to conclude bioequivalence.

12.2.2 Design of New Bioequivalence Study

By the time we began to design a new bioequivalence trial, in addition to the original test formulation (T1), pharmaceutical formulation scientists had developed four other formulations (T2, T3, T4, T5) that management wanted to assess whether their bioavailability was comparable to that of the standard S. So, the protocol of the new bioequivalence trial included the standard S and all five test formulations.

12.2.3 Six-by-Six Latin Square Design Versus Six-Group Parallel Design

Initially, we considered using a six-by-six Latin square as the design, with each subject receiving each of the six formulations. We estimated sample size based on the first period data from the earlier two-by-two crossover study and concluded that we needed 20 subjects in each of the formulation groups to have 80% power (with a Type I error of 0.05) to detect a 20% difference in bioavailability endpoints for each pairwise comparison of a test formulation to the standard, if run as a six formulation group parallel study. This gave a total of 120 (between subjects) observations, 20 per group. We therefore were satisfied that running a six-by-six crossover study with four subjects per sequence (for a total of 24 subjects and 144 observations) would provide at least 80% power to detect a 20% difference between the formulations in each five pairwise comparisons of a test to the standard.

However, we estimated costs of conducting the crossover study versus conducting the parallel study. Obviously, based on the costs of concentration assays, the crossover

study was more expensive. The crossover was also expected to be more expensive in terms of volunteer stipends, since we expected that we would have to pay volunteers 'pretty big bucks' to ensure that they participated in all six periods. In addition, facility rental costs for the crossover study (6 weekends) would be more expensive than for the parallel study (5 weekends with 24 subjects each). So, the final study design was a six formulation, parallel group study with 20 subjects per group.

12.2.4 Sample Size Re-estimation

As the variance estimate for sample size estimation derived from the first period data from the referenced two-by-two crossover study with only eight subjects per sequence, we decided to include a sample size re-estimation plan. Twenty-four different subjects participated in the trial on each of 5 consecutive weekends. During each weekend, subjects were given the formulation to which they were randomized and blood samples taken at times specified in the protocol. Concentrations of drug in the samples became available prior to entering the next group of 24 subjects.

Our data management department developed a computer program which 'kicked out' in blinded fashion only the variance estimate based on the analysis model after the data (AUC) were available from the subjects entered on each weekend (before the next weekend). After the second weekend (48 subjects, 8 per group) the variance estimate used for sample size estimation was greater than the variance estimate based on the 48 subjects. We thus concluded that our sample size estimate of 120 was adequate.

It is noted that the sample size re-estimation plan introduced no bias, as under normal theory the estimators of variance and mean are independent (so that knowledge of the variance estimate as data accumulated gave us no information about mean differences between formulations), neither did the sample size re-estimation plan invoke a Type I error penalty (as no between formulation groups inference was made).

12.3 A 2 × 2 Factorial Design in Assessing Efficacy of a Fixed Combination Product in Seasonal Allergic Rhinitis (SAR) (Peace and Chen 2010; Diamond et al. 1981; Peace and Tsai 2009; Peace 2005)

12.3.1 Background

A fixed combination product (TP) of triprolidine (T) and pseudoephedrine (P) was marketed prior to the DESI review in the 1970s. The DESI review deemed that a

clinical trial would have to be conducted to demonstrate the effectiveness of TP for this indication.

12.3.2 Design

The experimental design of the trial incorporated a 2×2 full factorial design with fixed, parallel, treatment groups: the combination (TP) product, triprolidine alone (T), pseudoephedrine alone (P), and placebo (0). Approximately, 160 patients with seasonal allergic rhinitis (SAR) were randomized in balanced fashion (40 per group) to the four treatment groups. Each patient received three doses of the drug to which he/she was randomized. The first dose was taken just after randomization at baseline, the second dose was taken 3 h after the first dose, and the third dose was taken 3 h after the second dose.

12.3.3 Objective

The objective of the trial was to demonstrate the efficacy of the combination product (TP) in the treatment of seasonal allergic rhinitis (SAR).

12.3.4 Primary Efficacy Endpoints

Hallmark signs and symptoms of SAR are nasal airway congestion (or resistance) and sneezing, rhinorrhea, lacrimation or itching of the eyes, nose, or throat. Therefore, primary efficacy endpoints were nasal airway resistance (NAR), ranging from rating of 1 to a rating of 6 in order of increasing severity, and a hay fever symptom complex score (HFSC) (occurrence and frequency of sneezing, rhinorrhea, lacrimation or itching of the eyes, nose, or throat), ranging from a score of 0 to a score of 44. These endpoints were assessed at baseline and hourly after each dose.

12.3.5 Objective Revisited

The fixed combination FDA regulation indicates that to prove that TP is effective, TP must be proven to be better than T and TP must be proven to be better than P. For a disease condition for which the primary efficacy measure is expected to be affected by both components of the combination, this requirement may be interpreted as the alternative hypothesis H_a in the compound hypothesis testing framework:

$$H_o : H_{o1} \cup H_{o2} \text{ versus } H_a : H_{a1} \wedge H_{a2}$$

where the individual null (H_{o1} and H_{o2}) and alternative hypotheses (H_{a1} and H_{a2}) appear in:

$$H_{o1} : \mu_{TP} = \mu_T \text{ versus } H_{a1} : \mu_{TP} > \mu_T, H_{o1} : \mu_{TP} = \mu_P \text{ versus } H_{a1} : \mu_{TP} > \mu_P,$$

where μ_{TP}, μ_T, μ_P, and μ_0 denote the true effect of the drug among patients in the respective groups in terms of the efficacy measure.

As T (antihistamine) is not expected to affect NAR and P (a sympathomimetic) is not expected to affect HFSCS, the objective, reflecting the fixed combination FDA regulation, is addressed by the following compound alternative hypotheses:

$$\text{For NAR, } TP > T \text{ and } TP > 0 \text{ and for HFSCS, } TP > P \text{ and } TP > 0.$$

The logic being by combining T and P, the efficacy of P as a decongestant must not be lost and the efficacy of T in relieving the signs and symptoms of hay fever must not be lost. Each was tested at the 0.025 level of significance.

In analyzing data from the trial, the primary efficacy measures had to be transformed (due to lack of normality). The logarithm of NAR and the logarithm of (HFSCS + 1) were statistically analyzed to address the trial objective.

Figures 12.1 and 12.2 illustrate means of Log(NAR) and Log(HFSCS + 1) by treatment group and post-baseline hour, respectively. Figure 12.3 represents a bivariate plot of the means of Log(NAR) and Log(HFSCS + 1) as a function of treatment group and hour of assessment. Baseline hour (0) and the hour after baseline (1, 2, …, 8) at which measurements were made appear along the graph of each treatment group. Graphs for each treatment group illustrate the effect of the drug in each group on both primary measures jointly. Movement of the graph for a treatment group in the direction of the origin reflects improvement.

12.4 A Dose Comparison Trial in Duodenal Ulcer (Peace and Chen 2010; Peace 2005; Peace et al. 1985; Dickson et al. 1985; Venezuela et al. 1985; Peace 2011)

12.4.1 Background

In the mid 1980s, based on data from gastric acid anti-secretory studies at various doses and frequencies of dosing, there was reason to believe that a single night time (hs) dose of 800 mg of the first H2-receptor antagonist cimetidine (C) for up to 4 weeks would be clinically optimal in treating duodenal ulcer patients. When first consulted, the original clinical development plan consisted of two, randomized, double-blind, placebo controlled, pivotal proof of efficacy trials. One trial would

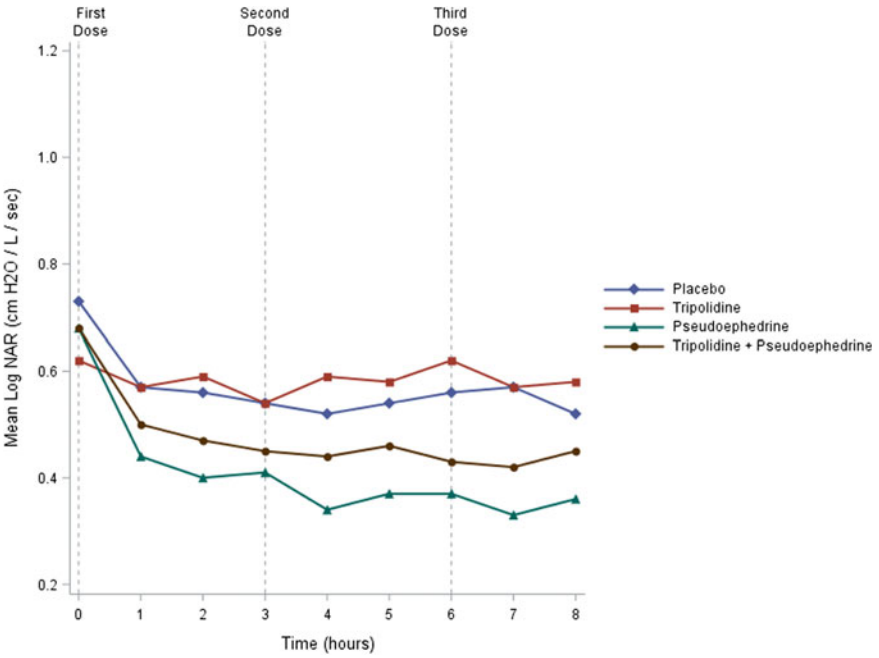


Fig. 12.1 Plot of mean log (NAR) by treatment group and time of assessment. Baseline values are shown at hour 0

compare 800 mg hs to placebo, while the other would compare 1200 mg hs to placebo.

Thus, the objective for each trial is the alternative hypothesis of the two constructs:

$$\text{Trial 1} : \text{Ho} : \mu_{800C} = \mu_P \text{ versus Ha} : \mu_{800C} > \mu_P$$
$$\text{Trial 2} : \text{Ho} : \mu_{1200C} = \mu_P \text{ versus Ha} : \mu_{1200C} > \mu_P$$

Each trial was to enroll 150 patients per treatment group, for a total of 600 patients. One hundred and fifty patients per group would provide a power of 95% to detect a 20% difference in cumulative 4-week ulcer healing rates between the C and placebo groups with a one-sided, Type I error of 5%.

Since conducting these two trials would subject ½ the patients to placebo and provide no in-trial comparison between dose groups (necessary to conclude optimality in some sense of the 800 mg hs regimen), the author recommended amalgamating the two trials into a single trial with three distinct treatment groups as per:

$$\text{Trial 3} : \text{1200 mg C hs versus 800 mg C hs versus 0 mg C hs (Placebo)}$$

with 164 patients per treatment group, for a total of 492 patients. One hundred and sixty-four patients per treatment group would provide a power of 95% to detect a

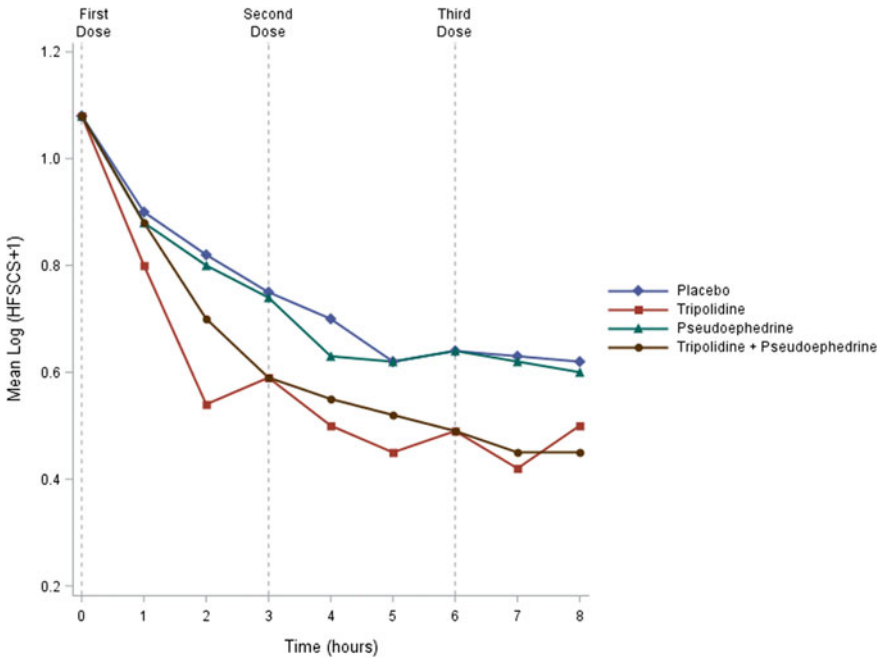


Fig. 12.2 Plot of mean log (HFSCS + 1) by treatment group and time of assessment. Baseline values are shown at hour 0

difference of 20% in cumulative 4-week ulcer healing rates between any two of the treatment groups with an experiment-wise Type I error of 5% (1.67% per each one-sided, pairwise comparison). Not only would this trial require fewer patients and be less expensive to conduct, it would also provide a within-trial comparison between C doses, for dose discrimination.

Further savings could be realized by incorporating into the Trial #3 protocol, a planned interim analysis after ½ the patients had been entered and completed. At the interim analysis, the efficacy comparisons: 1200 mg C vs. placebo and 800 mg C vs. placebo would be tested. If both were statistically significant, then the entire study could be stopped providing efficacy of each C dose were the only objective. If comparing the doses of C was also of clinical importance, then the placebo arm could be stopped and the two C arms run to full completion to have greater power in the comparison of doses. By conducting Trial #3 (instead of the two separate trials) and incorporating the interim analysis, potential savings of up to 190 patients could be realized, a savings of approximately $2 million. Additional savings would be expected due to less time required to conduct the trial.

However, Trial #3 was not conducted. Instead, a landmark dose comparison trial was conducted retaining the 800 mg hs dose and placebo, but including a dose lower than 800 mg hs and a dose higher than 1200 mg HS. The primary objective in conducting a clinical trial of C in the treatment of duodenal ulcers with a single

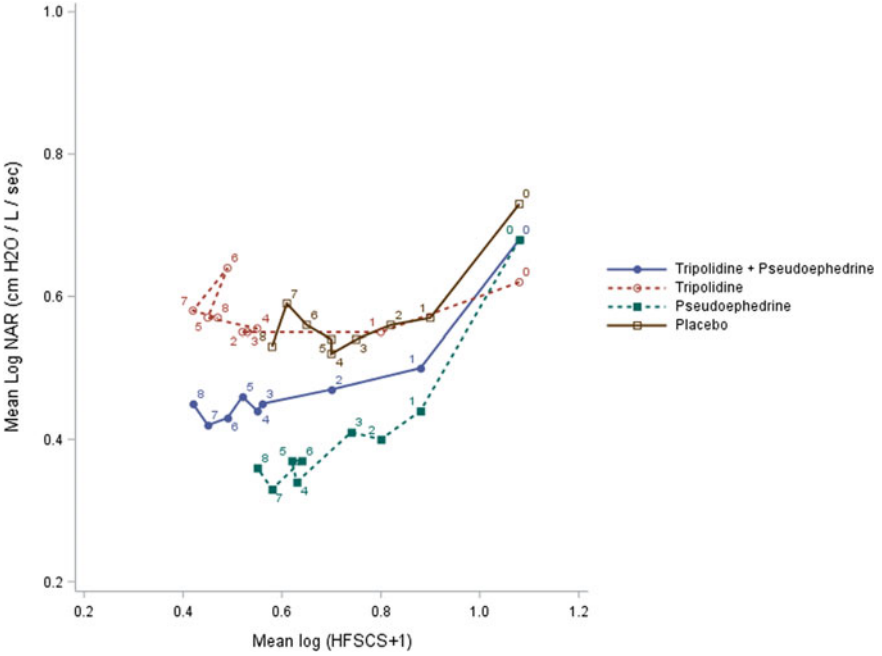


Fig. 12.3 Plot of mean log (NAR) and mean log (HFSCS + 1) by group and hour of assessment. Baseline values are at hour 0

nighttime dose was to demonstrate that 800 mg C was clinically optimal. We therefore added a 400 mg dose and replaced the 1200 mg dose with a 1600 mg dose (a twofold increase among consecutive doses) in the final trial protocol, which was IRB and FDA approved.

12.4.2 Landmark Dose Comparison Trial

Both primary and secondary efficacy objectives were identified in the final protocol. The primary objective addressed ulcer healing. The secondary objective addressed upper gastrointestinal (UGI) pain relief.

12.4.2.1 Primary and Secondary Objectives

The primary objective was to confirm that C given as a single nighttime dose of 800 mg for up to 4 weeks was clinically optimal in healing duodenal ulcers. Clinically optimal meant that 800 mg C was effective (significantly superior to placebo), that 800 mg C was superior to 400 mg C, and that 1600 mg C was not significantly

superior to 800 mg C. Symbolically, the primary (note p subscript of H) objective derives from three null and alternative hypotheses:

$$H_{p01} : P_{uh800} = P_{uh0} \cdots\cdots\cdots H_{p02} : P_{uh800} = P_{uh400} \cdots\cdots\cdots H_{p03} : P_{uh1600} \neq P_{uh800}$$
$$H_{pa1} : P_{uh800} > P_{uh0} \cdots\cdots\cdots H_{pa2} : P_{uh800} > P_{uh400} \cdots\cdots\cdots H_{pa3} : P_{uh1600} = P_{uh800}$$

where P_{uh0}, $P_{uh400,}$ P_{uh800}, and P_{uh1600} represent the cumulative ulcer healing (uh) rates by week 4 in the placebo, 400 mg C, 800 mg C, and 1600 mg C treatment groups, respectively, under single nighttime (hs) dosing. Specifically, the primary objective is the compound alternative H_{pa} of the hypothesis testing construct: H_{p0}: H_{p01}U H_{p02} U H_{p03} versus H_{pa}: $H_{pa1} \wedge H_{pa2} \wedge H_{pa3}$ comprised the primary study objective. Addressing the objective by individual analysis of each of the three univariate hypotheses, it is clear that the experiment-wise Type I error rate of 0.05 must be preserved across the three separate univariate hypotheses.

Symbolically, the secondary (note s subscript of H) objective derives from the three null and alternative hypotheses:

$$H_{s01} : P_{pr800} = P_{pr0} \cdots\cdots\cdots H_{s02} : P_{pr800} = P_{pr400} \cdots\cdots\cdots H_{s03} : P_{pr1600} \neq P_{pr800}$$
$$H_{sa1} : P_{pr800} > P_{pr0} \cdots\cdots\cdots H_{sa2} : P_{pr800} > P_{pr400} \cdots\cdots\cdots H_{sa3} : P_{pr1600} = P_{pr800}$$

where P_{pr0}, $P_{pr400,}$ P_{pr800}, and P_{pr1600} represent the UGI pain relief (pr) rates in the placebo, 400 mg C, 800 mg C, and 1600 mg C treatment groups, respectively, under single nighttime (hs) dosing. Specifically, H_{sa1}, H_{sa2}, and H_{s03} comprised the secondary study objective.

Of the six possible pairwise comparisons among the four dose groups, only three comprised the study objective. The other three: 1600 mg C versus 0 mg C, 1600 mg C versus 400 mg C, and 400 mg C versus 0 mg C were not part of the study objective and thus did not exact a Type I error penalty (i.e., the overall Type I error of 5% was 'Bonferonnied' across the three pairwise comparisons comprising the study objective, and not across the six possible pairwise comparisons).

12.4.2.2 Sample Size Determination

The trial was designed to enter enough patients to complete one hundred and sixty-four (164) per treatment group, for a total of 656 patients. One hundred and sixty-four patients per treatment group would provide a power of 95% to detect a difference of 20% in cumulative 4-week ulcer healing rates between any two of the treatment groups with an experiment-wise Type I error rate of 5% (1.67% for each of the three, one-sided, pairwise comparisons, reflecting the study objective). The worst case of the binomial variance was used in estimating the sample size since prior duodenal ulcer clinical trials of cimetidine showed an approximate 50% healing rate in the placebo groups

12.4.2.3 Further Design Considerations

The trial was multicenter, stratified, randomized, double-blind, and placebo (0 mg C) controlled. Neither patients, investigators, nor their staff knew the identity of the four treatment regimens. As there had been reports that there might be a negative correlation between smoking status and duodenal ulcer healing, patients were stratified by heavy smoking status (Yes, No) within each center prior to randomization to the four treatment groups.

In addition, since there was pressure to conduct the trial as quickly as possible, we planned to enlist a large number (approximately 70) of investigational sites. So, we expected there would be many centers who enrolled very few patients and therefore would impact the analysis and interpretation using the completely randomized design block (CRBD) model, with investigational sites as blocks. Since we believed that there would also be a negative correlation between ulcer size and ulcer healing, we defined six ulcer size categories: [0.3], (0.3; 0.4], (0.4; 0.5], (0.5; 1.0), [1.0], and (1.0; 3.0]. Our primary analysis model would not include investigational site as a block, but would instead include smoking status-by-ulcer size as blocks (12 blocks, where the 11 total degrees of freedom partition into 1 for smoking, 5 of ulcer size, and 5 for interaction between smoking and ulcer size).

12.4.2.4 Blinded Treatment Groups

Blinded treatment group medication was packaged using the existing regulatory approved 400 mg C tablet. A 400 mg placebo tablet was formulated identical to the 400 mg C tablet except that it contained 0 mg C. Blinded trial medication for the four treatment groups was packaged in blister packs for 4 weeks of nightly treatment as:

0 mg C Group: Four 400 mg placebo tablets;
400 mg C Group: One C 400 mg tablet + three 400 mg placebo tablets;
800 mg C Group: Two C 400 mg tablets + two 400 mg placebo tablets; and
1600 mg C Group: Four C 400 mg tablets.

12.4.2.5 Entry and Assessment Procedures

To enter the trial, patients were required to have an endoscopically confirmed duodenal ulcer of size at least 0.3 cm, and either daytime or nighttime UGI pain. After providing informed consent at the baseline visit, patients provided a history (including prior use of medications, particularly anti-ulcer ones or antacids), underwent a physical examination, had vital signs measured, provided blood and urine samples for clinical laboratory assessments, had UGI pain assessed, and underwent endoscopy. Patients were instructed how to use a daily diary to record the severity of daytime or nighttime UGI pain, as well as to record any adverse experience or concomitant medication use. Diaries and trial medication were dispensed, and

the patients were instructed to return at weeks 1, 2, and 4 of the treatment period for follow-up endoscopy, UGI pain assessment, and assessment of other clinical parameters. Antacids were provided to patients for relief of severe pain during the first 6 days/nights of therapy only and were limited to four tablets per day of low acid-neutralizing capacity.

Follow-up endoscopic evaluation was carried out following strict time windows at week 1 (days 7–8), week 2 (days 13–15), and week 4 (days 26–30). Patients whose ulcers were healed at any follow-up endoscopy were considered trial completers and received no further treatment or endoscopic assessment.

12.4.3 Innovative Aspects of the Clinical Trial Program

There are several aspects of this program that were rather innovative and 'firsts.'

12.4.3.1 Interim Analyses to Drop Placebo Arms

Interim analyses plans that would allow dropping of the placebo arm after establishing efficacy of the doses, while allowing the dose arms to run to completion for dose discrimination, were developed.

12.4.3.2 Trial Objectives as Only Three of Six Pairwise Comparisons

The study objective was formulated as only three of six pairwise comparisons among the four dose groups while preserving the overall experiment-wise Type I error across these three comparisons. The other three comparisons could be investigated, preferably using confidence intervals, but they should not invoke a Type I error penalty on the study objective.

12.4.3.3 Giving up Information on Center Differences

Instead of using centers as a blocking factor in the primary analyses, the 12 classifications of smoking status-by-baseline ulcer size was used as the blocking factor due to small numbers of patients per treatment group per center and due to the prognostic importance of smoking status and baseline ulcer size.

12.4.3.4 Assessment of Type of Monitoring Group

Roughly, half the investigational sites were recruited and monitored by in-house clinical operations personnel. The remaining half were recruited and monitored by

an outside Contract Research Organization (CRO). An assessment of differences in treatment effect between sites monitored by in-house personnel and those monitored by the CRO was conducted. There was no treatment-by-type of monitoring interaction, although the healing rates were generally lower among CRO monitored sites.

12.4.3.5 Association Between Ulcer Healing and Smoking Status and Ulcer Size

This landmark duodenal ulcer trial definitively established for the first-time negative correlations between ulcer healing and smoking status and ulcer healing and baseline ulcer size. Smokers experienced a significant lower healing rate than nonsmokers. Patients with the largest ulcers were less likely to heal than patients with smaller ulcers. Smokers with largest ulcers were most difficult to heal, whereas nonsmokers with smallest ulcers were easiest to heal. Effectiveness estimates of ulcer healing were adjusted for smoking status and baseline ulcer size.

12.4.3.6 Bivariate Graphical Methods

The duodenal ulcer trial was the first to utilize bivariate plots to profile ulcer healing and UGI pain relief jointly. The plots illustrated strong dose response in terms of ulcer healing and UGI pain relief separately and jointly (Fig. 12.4). It is noted in Fig. 12.4 that not all patients had daytime pain at baseline (not all treatment group graphs begin at one on pain axis). The reason for this is that the protocol called for patients to have ulcer-like pain during the day **or** during the night.

12.4.3.7 Establishing Effectiveness Based on a Subset Analysis

This trial was conducted at a time when the FDA began to let it be known that they expected the NDA or SNDA to address whether the effect of treatment seen in the entire clinical trial efficacy population generalized over subsets (by sex, by race or ethnicity, by age, by disease severity) of the population. We established from the trial that the 800 mg C dose was effective in the elderly population based on a subset analysis of patients 65 years old or older. In addition, labeling included acknowledging that for patients who smoked and had large ulcers, 1600 mg hs could be used.

12.4.3.8 Maximum Use of Patients with UGI Pain Who Were Screened

The landmark, dose comparison trial of once nightly C in the treatment of duodenal ulcer is but one of three clinical trials we conducted simultaneously at the investi-

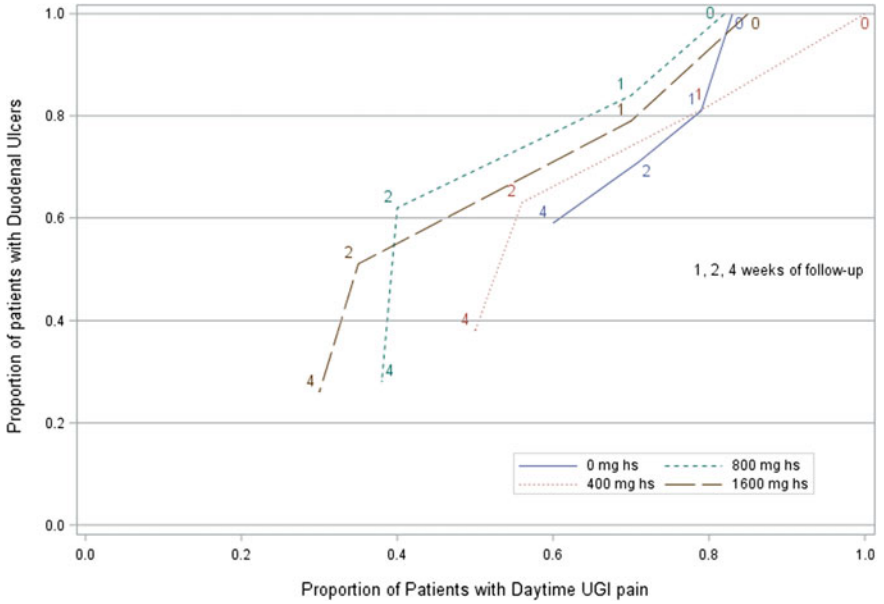


Fig. 12.4 Proportions of patients with unhealed ulcers and proportion of patients with daytime pain by dose group and week of endoscopy

gational sites. Each center conducted three protocols: the one discussed in duodenal ulcer patients, one in gastric ulcer patients, and one in patients with dyspepsia. This represented maximal use of patients who were screened.

To expand further, patients were recruited if they experienced ulcer-like symptoms including epigastric UGI pain. Those who satisfied general entry criteria and who gave consent underwent endoscopy. If duodenal ulcer (DU) and not gastric ulcer (GU) was confirmed, they entered the landmark DU trial. If GU was confirmed, they entered a GU trial, and if there was no DU and no GU, they entered a dyspepsia trial. This latter protocol provided a rather stringent definition of dyspepsia: Ulcer-like symptoms including epigastric UGI pain not explained by the presence of DU or GU upon endoscopy. This concurrent protocol method maximized the utility of the advertisement effort to get patients to the clinic who were experiencing ulcer-like symptoms.

12.4.4 Conclusions

The SNDA clinical trial program that led to approval of clinically optimal dosing of the first H2-receptor antagonist: Cimetidine in the treatment of duodenal ulcers has been reviewed. The program included a landmark clinical trial that not only

definitively established 800 mg C hs for 4 weeks as the clinically optimal dosing regimen, but also was the first to definitively establish negative associations between ulcer healing and smoking status and ulcer size, as well as the first trial to establish bivariate dose response in terms of ulcer healing and relief of UGI pain. Clinical optimality of 800 mg C hs was defined as 800 mg C being effective as compared to placebo; 800 mg C being more effective than 400 mg C; and 1600 mg C not being more effective than 800 mg C.

Further, to make maximal use of patients screened, the program included clinical trials of the 800 mg C regimen in dyspepsia and in gastric ulcers. The program also included drug interaction trials of the 800 mg C dose with widely used drugs and a bioequivalence trial of a new 800 mg C tablet to be marketed compared to two, 400 mg tablets of the commercially available formulation. The bioequivalence trial was required as the clinical trial in DU was conducted using the commercially available 400 mg tablet at the time of study conduct.

12.5 Equiradial Hexagonal Design with Response Surface Methodology in Phase II Anti-anginal Trials (Peace and Chen 2010; Peace 1990)

12.5.1 Introduction

Angina pectoris is pain in the chest that occurs when the heart muscle receives blood that has reduced oxygen levels. The discomfort may also occur in other areas than the chest, for example: the back, shoulders, arms, neck, or jaw, and mimic the symptoms of indigestion. Angina is thought to be a symptom of coronary artery disease (CAD) rather than a disease per se. The first drug (coronary vasodilator) approved for the treatment of angina was nitroglycerin, administered under the tongue.

Response surface methodology (RSM) was incorporated into two Phase II clinical trials of an unapproved, in-licensed drug believed to have anti-anginal efficacy. The objective of the RSM was to estimate dose and frequency of dosing that could be used in developing Phase III, pivotal proof of anti-anginal efficacy protocols. The primary measure of anti-anginal efficacy in the Phase II protocols was time to onset of exercise-induced angina. An equiradial hexagonal design—two equilateral triangles with a common center point—and a full quadratic response model were used. Design and analysis aspects of these trials are reviewed in this chapter.

12.5.2 Original Objective of Phase II Protocols

The **original objective** of the Phase II protocols was to obtain dose comparison information on measures reflecting possible anti-anginal efficacy when given b.i.d.

and to compare the top b.i.d. regimen with a t.i.d. regimen at the same total daily dose. As the protocols were exploratory, Phase II studies and sample sizes were not determined from a power perspective, rather they were chosen to ensure replication in each treatment group (3) per center (2) per protocol (2). Essential features of the protocols are captured in the sections that follow.

12.5.3 Treatment Groups in the Original Protocols

The treatment groups for Protocol 1 were: 0 mg b.i.d., 4 mg b.i.d., 6 mg b.i.d., and 4 mg t.i.d. The first three regimens permit comparisons among doses: 0 (placebo), 4, and 6 mg when given b.i.d. The last two regimens permit the comparison of the b.i.d. and t.i.d. regimens at the same total daily dose (12 mg).

The treatment groups for Protocol 2 were: 2 mg b.i.d., 8 mg b.i.d., and '6 mg t.i.d.' The 2 mg b.i.d. regimen was thought to be a 'no effect' dose. The last two regimens permit the comparison of the t.i.d. regimen '6 mg t.i.d': 4 mg, 6 mg, and 6 mg, to the 8 mg b.i.d. regimen, at the same total daily dose (16 mg).

12.5.4 Efficacy Measures

For each protocol, measures of efficacy were: time-to-stress-test-induced anginal onset, total exercise time, double product (heart rate x systolic pressure), at the onset of angina and at the end of the exercise time, maximal *St* wave depression, time to maximal *St* wave depression, and weekly anginal frequency (with nitroglycerin use recorded). **Time-to-stress-test-induced anginal onset** was considered the **primary efficacy** measure.

12.5.5 Stress Testing and Dosing Considerations

Patients who were candidates for protocol entry underwent a stress test prior to randomization. Those who qualified for entry returned to the clinic a week later for a baseline stress test, randomization, and dispensing of study medication. They then returned to the clinic on day 17 for stress tests at 2, 7, or 12 h after the dose taken on day 17.

After randomization patients were dosed either b.i.d. or t.i.d. for 16 days, plus 0 or 1 tablet on day 17, dependent on the assigned dose group and timing of stress test on day 17.

Table 12.1 Coded and uncoded vertices of equilateral, hexagonal design

Protocol 1			Protocol 2		
Dose	Coded vertices	Uncoded vertices	Dose	Coded vertices	Uncoded vertices
4 t.i.d.	$\left(-1/2, -\sqrt{3}/2\right)$	(12 mg, 2 h)	2 b.i.d.	$\left(-1/2, -\sqrt{3}/2\right)$	(4 mg, 2 h)
6 b.i.d.	$\left(1/2, +\sqrt{3}/2\right)$	(12 mg, 12 h)	**2 b.i.d.**	$\left(-1/2, -\sqrt{3}/2\right)$	(4 mg, 12 h)
4 b.i.d.	(0, 0)	(8 mg, 7 h)	**4 b.i.d.**	(0, 0)	(8 mg, 7 h)
0 b.i.d.	(−1, 0)	(0 mg, 7 h)	6 t.i.d./8 b.i.d.	(1, 0)	(16 mg, 7 h)

12.5.6 Final Study Design to Permit Use of RSM

At the initial consultation, it was observed that by replicating the 2 mg b.i.d. group and adding a 4 mg b.i.d. group in Protocol 2 (see highlighted dose in Table 12.1), the two protocols could be amalgamated under a single equiradial hexagonal design. This design is a member of the class of uniform precision designs. It permits exploration of the efficacy measure using response surface methodology (RSM) as a function of total daily dose (0–16 mg) and time of stress test after the last dose (2, 7, or 12 h).

The design represents two equilateral triangles with a common center point (Fig. 12.5). The vertices of the hexagon are the vertices of the two equilateral triangles. The vertices have been transformed (Table 12.1) so that the center point (8 mg, 7 h) becomes (0, 0). The vertex (−1,0) represents 0 mg (placebo) total daily dose and stress test administered 7 h after the last dose (on day 17). The vertex (1, 0) represents 16 mg total daily dose (either 8 mg b.i.d. or '6 mg t.i.d.') and stress test administered 7 h after the last dose. The vertex $\left(-1/2, -\sqrt{3}/2\right)$ represents 4 mg total daily dose and stress test administered 2 h after the last dose. The vertex $\left(-1/2, -\sqrt{3}/2\right)$ represents 4 mg total daily dose and stress test administered 12 h after the last dose. The vertex $\left(-1/2, -\sqrt{3}/2\right)$ represents 12 mg total daily dose and stress test administered 2 h after the last dose. The vertex $\left(1/2, +\sqrt{3}/2\right)$ represents 12 mg total daily dose stress test administered 12 h after the last dose.

12.5.7 RSM Analysis Model

The full quadratic response surface model regressing time-to-delay in angina onset (TTDAO) is given by:

$$\text{TTDAO} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \xi,$$

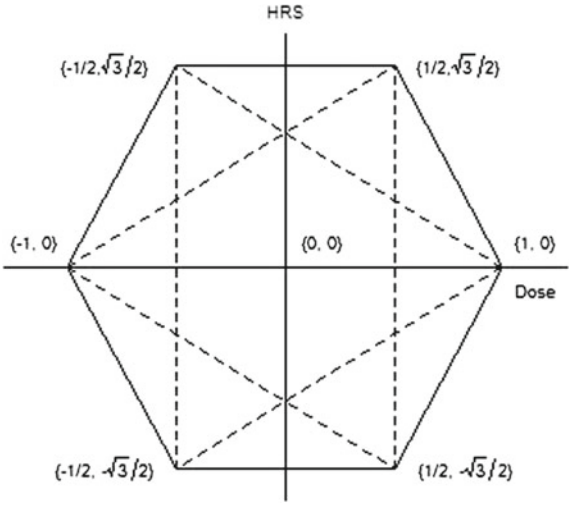


Fig. 12.5 Equiradial hexagonal design or two equilateral triangles with a common center point. (*Note* total daily dose and time of last stress test have been coded; see Table 12.1)

where X_1 represents total daily dose and X_2 represents the time of administering the stress test after the last dose.

TTDAO represented the difference between the time-to-anginal onset at the baseline stress test and the time-to-anginal onset stress test, at either 2, 7, or 12 h after the last dose.

Before settling on the full quadratic model given above, we included blocking variables to assess whether any differences between protocols and any differences between investigational sites within protocols were statistically significant. Seeing none, we settled on the full quadratic model given without including any blocking variables.

So, we fit the model to the data using PROC RSM of the statistical analysis system (SAS). The aim of these analyses was to identify total daily dose and time after last dose for which there existed acceptable clinical efficacy (defined a priori as delay in time-to-anginal onset at least 30% above baseline).

12.5.8 Analysis Results

Table 12.2 presents the analysis results from PROC RSM. Figure 12.6 shows the contour plot of the fitted response surface. Parenthetically, the fitted response surface must be searched to find regions in the X_1X_2 plane that correspond to expected clinical delay of anginal onset of at least 30% above baseline. We note that

(1) The model is not rejected due to lack of fit ($P > 0.52$) and that the model explains over 60% of the variation in response.

Table 12.2 ANOVA summary, RSM analysis

Response mean	1.271				
Root MSE	0.766				
R-square	0.531				
CV	60.308				
Regression	*DF*	*Type I SS*	*R-Square*	*F*	*Prob.*
Linear	2	23.287	0.442	19.82	<0.0001
Quadratic	2	2.016	0.038	1.72	0.1922
Cross product	1	2.667	0.051	4.54	0.0390
Total regress	5	27.969	0.531	9.52	<.0001
Residual	*DF*	*SS*	*MS*	*F*	*Prob.*
Lack of fit	1	0.250	0.250	0.42	0.5207
Pure error	41	24.420	0.596		
Total error	42	24.670	0.587		
Parameter	*DF*	*Estimate*	*SD*	*T*	*Prob.*
Intercept	1	1.450	0.221	6.55	<0.0001
X_1	1	0.950	0.181	5.26	<0.0001
X_2	1	−0.542	0.156	−3.46	0.0012
$X_1 * X_1$	1	0.050	0.313	0.16	0.8738
$X_1 * X_2$	1	−0.667	0.313	−2.13	0.0390
$X_2 * X_2$	1	−0.396	0.235	−1.69	0.0991
Factor	*DF*	*SS*	*MS*	*F*	*Prob.*
X_1	3	18.927	6.309	10.74	<0.0001
X_2	3	11.380	3.793	6.46	0.0011

(2) The model has predictive capacity; i.e., total regression is statistically significant ($P < 0.0001$).
(3) The significance of total regression is explained primarily by the significance of the linear terms ($P < 0.0001$) in the model.
(4) Both dose (X_1: $P < 0.0001$) and time of stress test after last dose (X_2: $P = 0.0012$) are statistically significant.
(5) The estimates of the coefficients (slopes) of dose and time of stress test after last dose are intuitively consistent; i.e., the predicted delay in time-to-anginal onset (P_DTTAO) from the model increases as dose increases (positive slope) and decreases as the time of last stress test after dose increases (negative slope).

Maximal predicted delay in time-to-anginal onset ranged from 2.98 to 3.16 min, representing a delay that ranged from 55 to 59% above baseline. However, the corresponding 'troughs' reflected unrealistic total daily dose and time of stress test after last dose, for example, a total daily dose of 16 mg and time of stress test 2 h after last

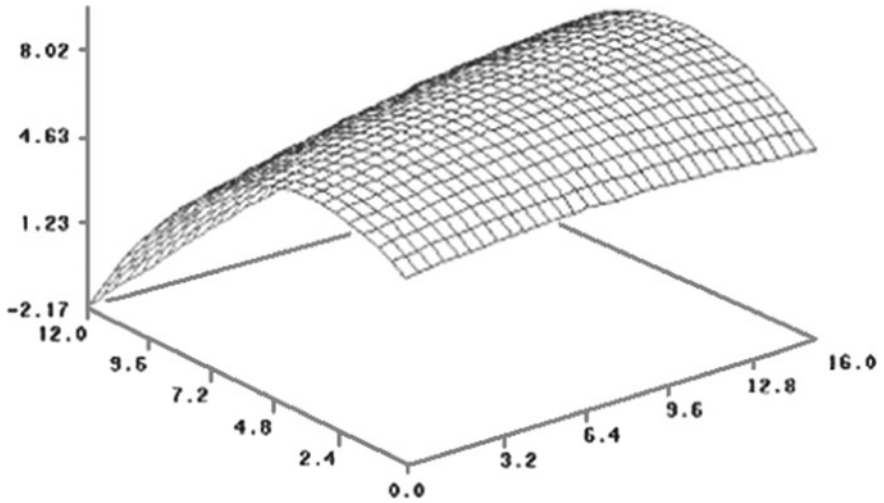


Fig. 12.6 Plot of fitted response surface. (*Note* total daily dose axis: 0.0–16.0; time of stress test after last dose 0.0–12.0 h)

Table 12.3 Summary of findings from RSM analysis

Desired effect	Predicted dose	Predicted time
P_DTTAO (min)	Increasing (mg)	Decreasing (h)
(2.98; 3.16) (55%; 59%)	16	2
(1.60;) (30%;)	8	2–5
(1.60;) (30%;)	16	8.5–9.5

dose—suggesting an unrealistic dose and frequency of dosing regimen of 1.33 mg given 12 times daily! (Table 12.3).

The specified clinically important delay in time-to-anginal onset of 30% is predicted with a total daily dose of 16 mg and frequency of dosing interval that contains 8 h. Therefore, the results of the RSM analyses suggest a total daily dose of 16 mg given t.i.d. (the 4, 6, and 6 mg regimen) for consideration in Phase III trials. Since the total daily dose was at the end of the dosing interval, it would be advisable for Phase III trials t.i.d. dosing regimens to bracket 16 mg per day, or to conduct another Phase II t.i.d. trial that brackets this dose.

References

Diamond, L., Gerson, K., Cato, A., Peace, K. E., & Perkins, J. G. (1981). An evaluation of triprolidine and pseudoephedrine in the treatment of Allergic Rhinitis. *Annals of Allergy, 47*(2), 87–91.

Dickson, B., Dixon, W., Peace, K. E., Putterman, K., & Young, M. D. (1985). Cimetidine single-dose active duodenal ulcer protocol design. *Post Graduate Medicine, 78*(8), 23–26.

Peace, K. E. (Editor and Author contributor). (1990). *Statistical issues in pharmaceutical drug development*. New York: Marcel Dekker, Inc. ISBN 0-8247-8290-9.

Peace, K. E. (2005). On the analysis of a combination of two primary efficacy measures. *The Philippine Statistician, 54*(4), 9–20.

Peace, K. E. (2011). Case study in optimal dosing in duodenal ulcer. *Invited submission: INTECH book on peptic Ulcer Disease* (ISBN 978-953-307-976-9).

Peace, K. E., Chen, D. (2010). *Clinical trial methodology*. Chapman & Hall/CRC, Taylor and Francis Group. ISBN 978-1-5848-8917-5.

Peace, K. E., & Tsai, K. T. (2009). Bivariate or composite plots of endpoints. *Journal of Biopharmaceutical Statistics, 19,* 324–331.

Peace, K. E., Dickson, B., Dixon, W., Putterman, K., & Young, M. D. (1985). a single nocturnal dose of cimetidine in active duodenal ulcer: Statistical considerations in the design, analysis and interpretation of a clinical trial. *Post Graduate Medicine, 78*(8), 27–33.

Venezuela, J., Dickson, B., Dixon, W., Peace, K. E., Putterman, K., & Young, M. D. (1985). efficacy of a single nocturnal dose of cimetidine in active Duodenal Ulcer: Result of a United States Multicenter Trial. *Post Graduate Medicine, 78*(8), 34–41.

Chapter 13
Uncovering Fraud, Misconduct, and Other Data Quality Issues in Clinical Trials

Richard C. Zink

13.1 Introduction

The quality of data from clinical trials has received a great deal of attention in recent years, as traditional approaches to assess quality through on-site monitoring, and 100% source data verification have come under increased scrutiny as providing little benefit for the substantial cost. Numerous regulatory guidance documents and industry position papers have described risk-based approaches, which makes use of central computerized review, to identify quality and safety issues (European Medicines Agency 2013; TransCelerate BioPharma 2013; US Food and Drug Administration 2013). An emphasis on risk-based approaches forces the sponsor to take a more proactive approach to quality through a well-defined protocol, sufficient training and communication, and by highlighting those data most important to patient safety and the integrity of the final study results. Further, identifying data problems early allows the sponsor to correct or refine study procedures as the trial is ongoing.

Unusual or problematic data can arise due to a number of reasons including carelessness (such as transcription errors), contamination of samples, mechanical failures, or miscalibrated equipment, poor planning (e.g., lack of appropriate backups or contingency planning should problems occur), or poor training in trial procedures. Fraud is also an important topic in discussions of data quality, and it distinguishes itself from the aforementioned issues due to the deliberate intention of the perpetrator to mislead others (Buyse et al. 1999). Google defines fraud as "wrongful or criminal deception intended to result in financial or personal gain." However, the difficulty or inability to discern the intention of, or the lack of any perceived benefit for, the

R. C. Zink (✉)
Data Management and Statistics, 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/10.1007/978-981-10-7820-0_13

culprit may make the term "misconduct," defined by Google as "unacceptable or improper behavior," more appropriate to use in practice.

Though the terms fraud and misconduct may have subtle differences in actual or perceived meaning, they are used interchangeably throughout this chapter to refer to data fabrication and falsification and to distinguish from other data quality issues that result from carelessness or any technical issues related to the generating or recording of data. Fabricating data occurs when fictitious data are recorded and submitted as if they were honestly obtained. This includes recording data for a patient for a procedure that may have been inadvertently missed, documenting study visits that did not take place, or making up patients in their entirety. Data falsification includes "manipulating research materials, equipment or processes, or changing or omitting data or results" so that data are not accurately reflected (Office of Research Integrity 2017). Examples of this include substituting the data of one subject for another, failing to record data that indicates a safety issue or altering data in order for patients to meet eligibility criteria and participate in the trial.

Despite the increasing availability of statistical and graphical tools available to identify unusual data, fraud itself is extremely difficult to diagnose (Buyse et al. 1999; Venet et al. 2012; Zink 2014). For one thing, many of the methods used to identify misconduct at a study center involve comparisons against other clinical trial sites. Such analyses could identify natural differences in patient population or variations in technique between the sites that would not constitute fraudulent behavior. Further, analyses motivated by a need to identify a particular type of malfeasance can detect data anomalies with reasonable explanations. Even if data turn out to be problematic, stating that the unusual findings are explicitly due to fraud may require evidence beyond what is available in the clinical trial database (Evans 2001).

Fraud in clinical trials is thought to be rare, though its prevalence is likely underestimated due to previously unavailable or limited tools and training for diagnosis, or for fear over negative publicity (Buyse et al. 1999; Weir and Murray 2011). Further, two recent publications describe higher than expected rates of scholarly retractions in the life science and biomedical literature, often due to fraud or suspected fraud (Fang et al. 2012; Grieneisen and Zhang 2012). Regardless of how extensive misconduct is in practice, recommendations to prevent clinical trial misconduct include simplifying study entry criteria, minimizing the amount of data collected, and utilizing sufficient and varied trial monitoring (Buyse et al. 1999; Baigent et al. 2008; Weir and Murray 2011; Venet et al. 2012). However, even in the presence of incorrect data due to manipulation or other quality issues, trial integrity will be preserved in most cases, most often due to randomization and blinding of study medication, or because the anomalies are limited to few sites (Buyse et al. 1999; Baigent et al. 2008; TransCelerate BioPharma 2013; US Food and Drug Administration 2013).

So this begs the question: If clinical trial fraud is so uncommon, with seemingly limited potential to seriously compromise the results of the trial, why bother looking for it at all? In short, it is much easier to identify problems as they occur while the trial is ongoing with the opportunity to resolve the issue or modify the trial as needed. Compare this to the scenario of finding a systemic problem once the trial has been unblinded and the final study results have been prepared. At this point, there are

fewer options available to the study team to find an appropriate solution, particularly when their every action will be scrutinized due to the availability of randomization codes. Most importantly, the early identification, resolution, and documentation of any lapse in data quality are important to protect the well-being of study participants (International Conference of Harmonisation 1996).

Clinical trial data are highly structured, and human beings are bad at fabricating realistic data, particularly in the many dimensions that would be required for it to appear plausible (Buyse et al. 1999; Evans 2001; Venet et al. 2012). Defining a series of statistical and graphical checks to be implemented on a regular basis to identify fraud is a minimal investment for the team to make to prevent potential catastrophe. Further, these same statistical checks can be used to identify data anomalies that occur in the absence of any misconduct. This chapter illustrates various statistical and graphical methodologies to assess data quality using a sample clinical trial. In Sect. 13.2, we summarize two examples of misconduct from the literature to motivate the regular screening of clinical trials for data quality issues. Section 13.3 briefly describes a clinical trial of patients who experienced an aneurysmal subarachnoid hemorrhage. This data will be analyzed in Sect. 13.4, with particular emphasis on the interpretation of analysis findings and suggestions for further analysis. Section 13.5 provides a brief conclusion. All analyses were performed using JMP Clinical 6.1.

13.2 Examples of Past Misconduct

13.2.1 Multicenter Animal Study

In the early 1980s, the National Heart, Lung, and Blood Institute conducted a multicenter animal study of two drugs with the ultimate goal to develop a reproducible animal model for studying myocardial infarction to evaluate new therapies (Bailey 1991). There were four centers where the research involved dogs. During data review, staff at the coordinating center identified some disparities between the data provided by one site when compared to the others. Study investigators were provided with data and were able to quickly determine that there were inconsistent relationships between the weights of the dogs and the weights of the left ventricle of the heart, as well as the infarct size and the level of blood flow to the heart from one of the centers. The project officer approached the laboratory chief of the site providing the unusual data, only to find out that the medical fellow responsible for conducting the experiments had fabricated data on a previous study.

In order to better understand the scope of the problem, the coordinating center set up an external panel of cardiovascular experts to further scrutinize the data. Additional review found noticeable differences between pre-discovery (Dogs 1–34) and post-discovery (Dogs 35 and above) dogs, but this was not viewed as sufficient to engage in formal proceedings of fraud. However, several pieces of hard evidence became available: Data were reported on a heart that was found to be discarded

without the planned studies having taken place, tissue from pre-discovery dog hearts had none of the expected radioactivity necessary to perform the assays, and data logs showed numerous discrepancies between the dogs reported and analyzed. Review of the medical fellow's past published work showed similar unusual data patterns to those identified in this study.

13.2.2 Dietary Intervention Trial

Allegations of fraud were made against the lead author of a manuscript describing the benefits of a dietary intervention in reducing heart attacks that was published in the British Medical Journal (BMJ) in 1992 (Al-Marzouki et al. 2005; White 2005). Suspicions were raised when the findings of a follow-up manuscript submitted to the BMJ later that year appeared too good to be true and inconsistent with the findings from previous clinical trials in the literature. Further, concerns over data collection, statistical methodology, and the similarity of this new work to other papers submitted by the lead author raised additional red flags. Because the lead author was outside the jurisdiction of an official research body or regulatory authority, the journal decided to investigate. Though the data for these two manuscripts were first requested in August of 1994, the final analysis was not available until March of 1999. Delays were initially due to the non-responsiveness on the part of the author, and subsequently due to the excessive amount of handwritten data that required entry. In the meantime, the author submitted and published manuscripts with other medical journals.

For the 1992 manuscript, re-analysis of the data identified numerous statistically significant differences among the means and variances of 22 covariates between the treatment and control arms at baseline, even after a Bonferroni adjustment for multiple comparisons. Further, 10 of 22 of the baseline characteristics exhibited significant differences among the trailing digit between the two groups (this methodology is described below). Based on these and other findings, the authors concluded that there was strong evidence of misconduct (Al-Marzouki et al. 2005). Additional investigation ensued, and lacking other alternatives, the BMJ went public with their findings in 2002.

13.3 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 at 40 clinical sites. Data were available for vital signs (heart rate, systolic, and diastolic blood pressure), 27 different laboratory tests, and 4 measurements from electrocardiogram (ECG) curves. Each day represented a different study visit.

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.

13.4 Screening Clinical Trials to Assess Data Integrity

13.4.1 Overview

The analyses below are inspired by the Buyse et al. (1999) paper where each site is, in turn, treated as the suspect site and compared to all other sites grouped together as a reference. This approach serves as straightforward means to implement a set of analyses quickly, particularly for groups with limited statistical support. Other approaches that consider the relationships between data collected from individuals within the same site, or sites within the same country are available (Desmet et al. 2013, 2017). Readers may also wish to consider how the multiple comparisons with the best (MCB) method of Hsu (1992) or the analysis of means (ANOM) method of Ott (1967) may be applied to analyses of data quality at clinical trial sites.

13.4.2 Multiplicity Considerations

Crowe and co-authors suggest using the false discovery rate (FDR) to adjust for multiplicity for safety outcomes in a clinical trial; we describe this approach in detail in Volume 2 of the *ICSA Biostatistics Book Series of the Biopharmaceutical Applied Statistics Symposium* (Crowe et al. 2009; Zink 2017). For the proactive screening of quality issues for data in a clinical trial, I sites with J procedures and K statistical tests to perform for each procedure can generate a large multiple testing problem. Appropriate multiplicity adjustment should achieve a reasonable balance between committing type I errors without overly sacrificing the power to detect potential quality signals, particularly when the study was not designed to detect lapses in quality between the study centers. The 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

Table 13.1 Trailing digit preference for a particular test

	0	1	2	3	4	5	6	7	8	9
Suspect										
Reference										

of multiple tests. In general, with H comparisons of ordered (smallest to largest) p values $p_{(h)}$, the FDR p value for the hth hypothesis is

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

Corresponding simultaneous 95% FDR confidence intervals can be defined by finding the largest h where $p_{(h)} \leq h\alpha/H$ and using $\alpha^* = h\alpha/H$ for all H confidence intervals (Benjamini and Yekutieli 2005). Below, we apply a separate $\alpha = 0.05$ to each analysis reported within Sect. 13.4.2 so that, at most, $H = I \times J$. However, alternate approaches could be used to adjust for multiple comparisons, such as a single $\alpha = 0.05$ across the $H = I \times J \times K$ tests. In cases where analyses are performed retrospectively due to suspected misconduct at a single trial site (as in Sect. 13.2.2), stronger type I error control through Bonferroni, or another method that preserves the familywise error rate, may be preferred (Westfall et al. 2011).

13.4.3 Sample Analyses

13.4.3.1 Analysis of Digit Preference

One example for identifying data anomalies involves analyzing digit preference for the data collected from the tests or procedures performed during a clinic visit, such as the vital signs, laboratory values, and ECG measurements described above. The distributions of leading or trailing digits (i.e., the first or last digit, respectively) can be compared between each suspect site and its corresponding reference. In this section, we focus solely on analyzing the trailing digit. A table similar to Table 13.1 would be populated with the frequencies n_{rc} (rth row and cth column) of observing each trailing digit for each suspect site and test combination, with as many as $H = (3 + 27 + 4) \times 40 = 1360$ tests performed for the sample data (assuming each test is observed at each site).

In lieu of a chi-square general association statistic, the Cochran–Mantel–Haenszel (CMH) row mean score statistic Q_S is used to take advantage of the ordinality of digit value for greater power; further, we apply standardized mid-rank scores to account for the possibility that the observed digits might not be equally spaced from one another

(Stokes et al. 2012). The row mean score statistic is $Q_s = \frac{(\bar{f}_1 - \bar{f}_2)^2}{\left\{\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right\}\left\{\frac{nv_a}{n-1}\right\}}$, where $v_a = \sum_{c=1}^{10} \frac{(a_c - \mu_a)^2 n_{+c}}{n}$, $\mu_a = \sum_{c=1}^{10} \frac{a_c n_{+c}}{n}$, and $\bar{f}_r = \sum_{c=1}^{10} \frac{a_c n_{rc}}{n_{r+}}$. Q_s is distributed $\chi^2_{(1)}$. Standardized mid-rank scores are defined as $a_c = \frac{2\left[\sum_{k=1}^{c} n_{+k}\right] - n_{+c} + 1}{2(n+1)}$.

The clinical and statistical significance of a large number of tests can be easily summarized using a volcano plot to highlight the combinations of greater interest: those tests with large numerical differences and/or those meeting the criteria for statistical significance after applying a suitable multiplicity adjustment (Zink et al. 2013, Zink 2014). In Fig. 13.1, 1288 site-test combinations are summarized. The x-axis represents the maximum difference between the suspect site versus the reference across all observed digits to represent the max percent difference. The y-axis represents the raw p value from the CMH row mean score test on the negative log10 scale, so the smaller the p value, the larger the value vertically. In general, the interesting site-test combinations are those that approach the upper corners and are above the dashed FDR reference line. Systolic and diastolic blood pressures for Sites 16 and 40 are labeled as points of interest to examine further.

Further analysis of the Site 40 markers provides the trailing digit bar charts in Fig. 13.2. Notice the investigator(s) at Site 40 were twice as likely to report a "0" in the trailing digit for both diastolic (left) and systolic (right) blood pressures, with "5" as the second most common result. How could we interpret these findings? Perhaps Site 40 may not be following the study protocol. For example, the investigators at that center may have collected blood pressures manually, reading values from a gauge where the best one could do is to interpret a 0 or 5 as the last digit, instead of using a machine to obtain more accurate measurements. Alternatively, this investigator might have a tendency, compared to the other sites, to round measurements to a 0 or 5. Further, investigation would be required to understand the reason behind such a finding. However, the study team first needs to decide if such a finding is substantial or important enough based on the study endpoints and goals to intervene.

Analyses of trailing digit preference can identify other issues aside from rounding—they can be used to detect instances where diagnostic equipment might be miscalibrated, where an error code may be output repeatedly, or the readings may be consistently too high or too low. Important differences in subjective measurements can be identified, such as the investigator's assessment of clinical signs using a Likert scale, which might suggest that additional training is needed. Analyses of leading digits can also identify sites that tend to read much higher or lower than the other sites. Alternatively, leading digits can be assessed to identify departures from Benford's Law, where each digit d is expected to be observed with probability $\log_{10}\left(1 + \frac{1}{d}\right)$ (Benford 1938; Hill 1996). The corresponding test is $\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(k-1)}$. However, as Kirkwood et al. (2013) noted in their analysis of digit preference, the comparisons of observed digits between sites are likely to be more useful since Benford's law may not apply in practice. Finally, while this section focused on the analysis of the trailing (or leading) digit for any numeric data, similar analyses could be used to compare the distributions of nominal or ordinal categorical variables between each suspect site and its reference. In the case of nominal categorical variables, how-

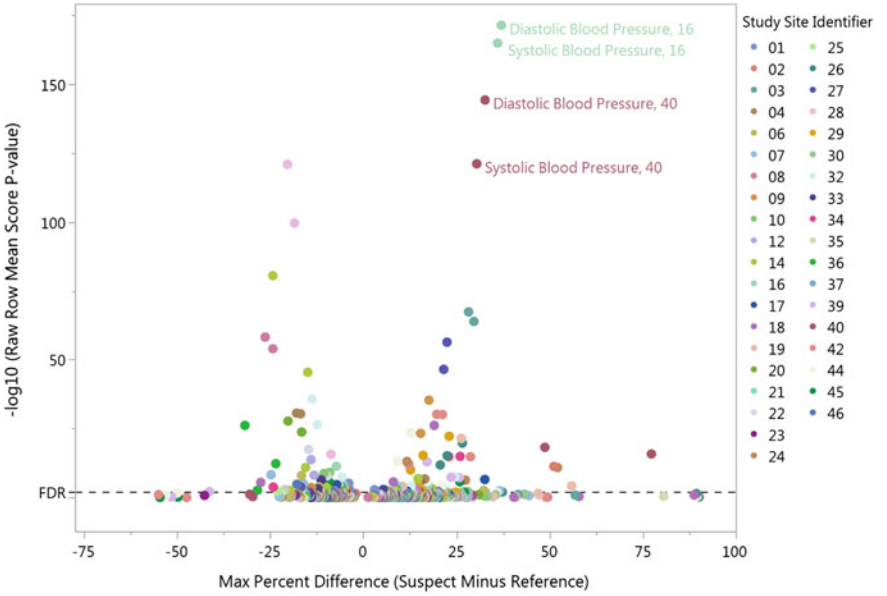


Fig. 13.1 Volcano plot of trailing digit preference among ECG, vital sign, and laboratory measurements (FDR) reference line drawn at $-\log10(0.00493) = 2.3071$, where $\alpha^* = 127/1288 \times 0.05 = 0.00493$. Alternatively, the FDR reference line could be drawn at $-\log10$ (maximum unadjusted p value $\leq \alpha^*$) as in Zink et al. (2013). Bubbles are colored according to clinical trial site. Each point represents a specific test, such as systolic blood pressure, for a specific site comparing that suspect site to all other sites as a reference. The max percent difference is the largest difference observed among all digit values between the suspect site and reference, i.e., $\max\{p_{s,0} - p_{r,0}, \ldots, p_{s,9} - p_{r,9}\}$, where $p_{s,j}$ and $p_{r,j}$ are the proportion of tests with a trailing digit of j for the suspect and reference, respectively

ever, the CMH row mean score test should be replaced with the chi-square general association statistic.

13.4.3.2 Correlation Among Related Covariates

In Sect. 13.2.1, we described how inconsistent relationships between variables helped identify an instance of misconduct. As noted in the introduction, it is challenging to fabricate data in the many dimensions that would be required for it to appear plausible. Even in the absence of fraudulent behavior, examining higher-order moments such as the variance, skewness, and kurtosis of each variable, or the pairwise correlation among pairs of variables within each site can uncover a problem in data quality. In this section, we focus on the analysis of pairwise correlations. Analyses of correlations have been applied in the past to detect irregularities from questionnaires obtained in clinical trials (Taylor et al. 2002).

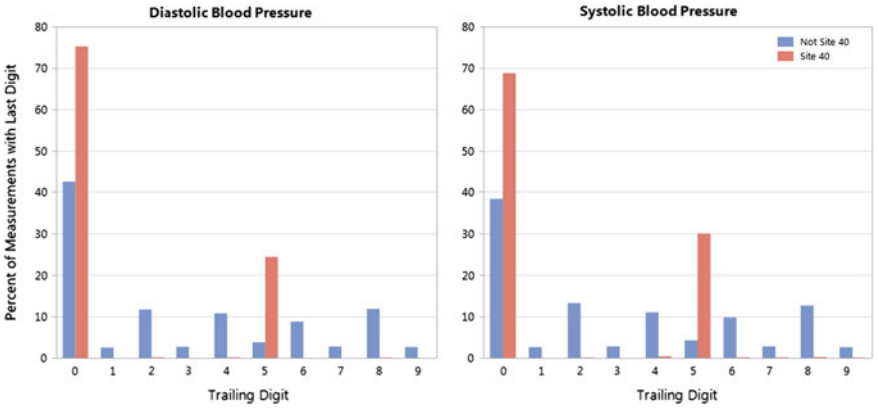


Fig. 13.2 Bar charts of trailing digit preference for diastolic and systolic blood pressure for site 40 compared to reference. The y-axis represents the proportion of all trial measurements that exhibit the trailing digit summarized along the x-axis

Fisher's Z-transformation can be used to calculate a test statistic to compare two correlation coefficients (Fisher 1921, 1970). Let $r_{ij_1j_2}$ be the correlation of variables j_1 and j_2 for site i. Fisher's Z-transformed correlation $r^*_{ij_1j_2}$ is defined as

$$r^*_{ij_1j_2} = \frac{1}{2}\log_e\left(\frac{1 + r_{ij_1j_2}}{1 - r_{ij_1j_2}}\right),$$

and the normally distributed test statistic comparing the correlation of variables j_1 and j_2 between groups s and r (suspect and reference, respectively) is defined as

$$Z_{srj_1j_2} = \frac{r^*_{sj_1j_2} - r^*_{rj_1j_2}}{\sqrt{\frac{1}{n_{sj_1j_2}-3} + \frac{1}{n_{rj_1j_2}-3}}}.$$

If variables are not approximately normal, it may be preferable to pre-transform according to a log transformation. Alternatively, Spearman's correlation, which is based on the ranks of the values, may be used. In general, I sites with J procedures performed will result in up to $\frac{I \times J \times (J-1)}{2}$ comparisons. However, since tests within each domain (vital signs, laboratories, ECG) tend to be measured at the same time (which may not be true across domains), we limit analyses to pairs of tests within the same domain. For this study, this would result in as many as $40 \times \left(\binom{3}{2} + \binom{27}{2} + \binom{4}{2}\right) = 14{,}400$ comparisons if all pairs of tests within domains were available in each site, with at least four instances of each measurement. Using the sample data, 13,680 tests were performed.

In Fig. 13.3, the x-axis represents the difference in untransformed correlations between the suspect site versus its reference. The y-axis represents the raw p value

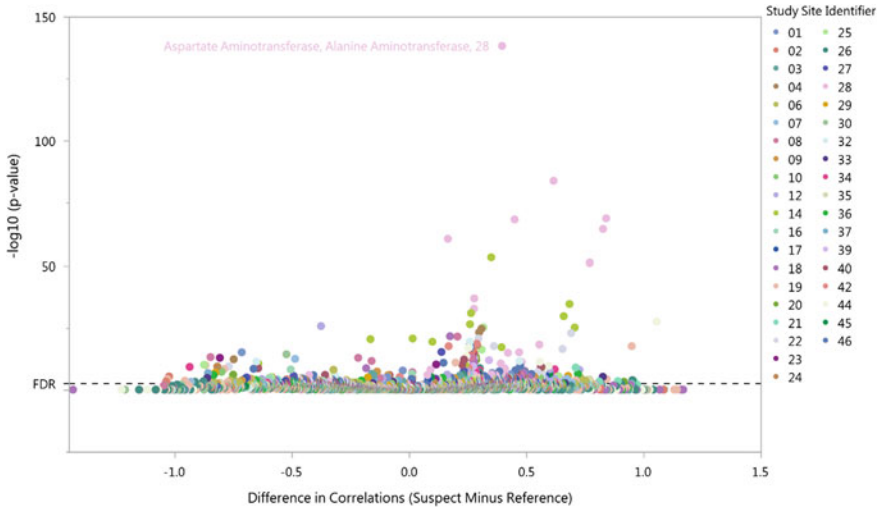


Fig. 13.3 Volcano plot of pairwise correlations among ECG, vital sign, and laboratory measurements. Pairwise correlations were only evaluated within domain (ECG, vital signs, or laboratories) when measurements were likely to be observed at the same time. The y-axis represents the p value for a test of the difference of Fisher's Z-transformed Pearson correlation coefficients. The x-axis represents the difference in Pearson correlation coefficients. FDR reference line drawn at $-\log10(0.00232) = 2.635$, where $\alpha^* = 634/13680 \times 0.05 = 0.00232$. Bubbles are colored according to clinical trial site. Each point represents a specific pair of tests, such as aspartate and alanine aminotransferase, for a specific site comparing that suspect site to all other sites as a reference

from the Z-test defined above on the negative $\log_{10}$ scale, so the smaller the p value, the larger the value vertically. In general, the interesting site-pair combinations approach the upper corners and above the dashed FDR reference line. Here, numerous results are identified among six pairs of laboratory tests for Site 28.

Further analysis of the laboratory tests for Site 28 is presented in Fig. 13.4, where a graphical representation of the lower triangle of the correlation matrices for Site 28 and its reference highlight the magnitude of pairwise associations. Notice that Site 28 has much stronger positive correlations for the six selected pairs (outlined in Fig. 13.4), which comprise various combinations of four laboratory tests: alanine aminotransferase (ALT), aspartate aminotransferase (AST), lactate dehydrogenase, and prothrombin time. As an aside, some authors note that fabricated data tend to exhibit stronger correlation than non-fabricated data (Akhtar-Danesh and Dehghan-Kooshkghazi 2003).

Figure 13.5 displays a scatterplot matrix of the data for the four variables from Site 28. Numerous bivariate outliers contribute to the strong correlations of these variables within Site 28, many of which come from a single patient. While extreme values for any single laboratory test may indicate a potential safety concern, multiple tests that are jointly extreme may indicate something more severe, such as drug-induced liver

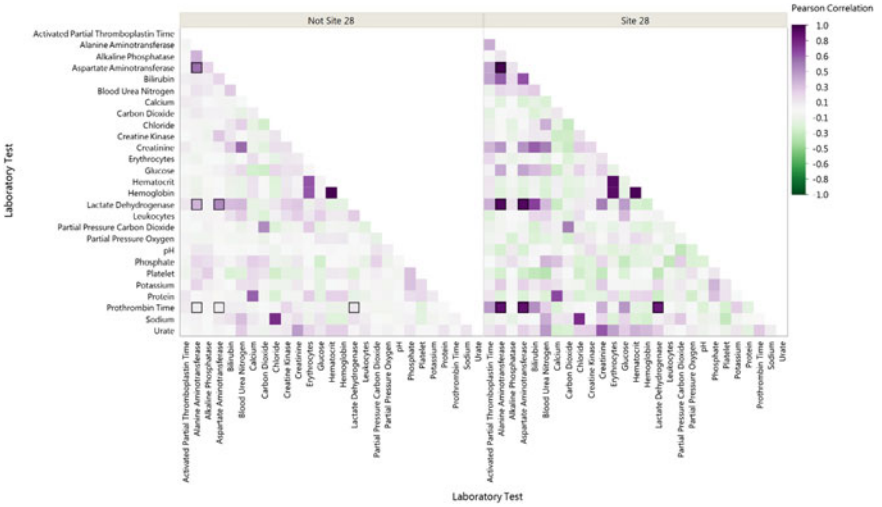


Fig. 13.4 Heat map of correlation matrix among laboratory measurements for site 28 compared to reference. Darker purple or green cells indicates pairs of tests with more positive or negative correlation, respectively

injury (US Food and Drug Administration 2009). Detecting such findings early can give the study team an opportunity to intervene on the patient's behalf.

The above analysis can be modified to assess the autocorrelation of repeated measurements over time, though it will likely be most useful when the repeated measurements are roughly evenly spaced. For example, each variable can be compared against a fixed lag with itself, say to assess the serial relationship of values that are one month apart. Noticeable differences in the autocorrelation between a suspect site and its reference could warrant further review. In extreme cases, it may be worthwhile to examine instances where the values for a particular test do not change at all over the course of the entire study, though this may occur in practice if limits of detection are consistently exceeded.

13.4.3.3 Visit Scheduling

In this final example, we examine the scheduling of study visits. Buyse and co-authors (1999) describe an example where study visits for a suspect site appeared too good to be true (Fig. 13.6). The study days on which the visit occurred were rather evenly distributed from Day 18 to 25 among the non-suspect sites. However, the suspect site had most visits occurring on Day 21, which is likely the expected day for the visit to take place. However, while this finding is unusual, it isn't necessarily the result of misconduct. Some investigators tend to block enroll and go out of their way to schedule trial patients on a limited number of days. So while the scheduling of this

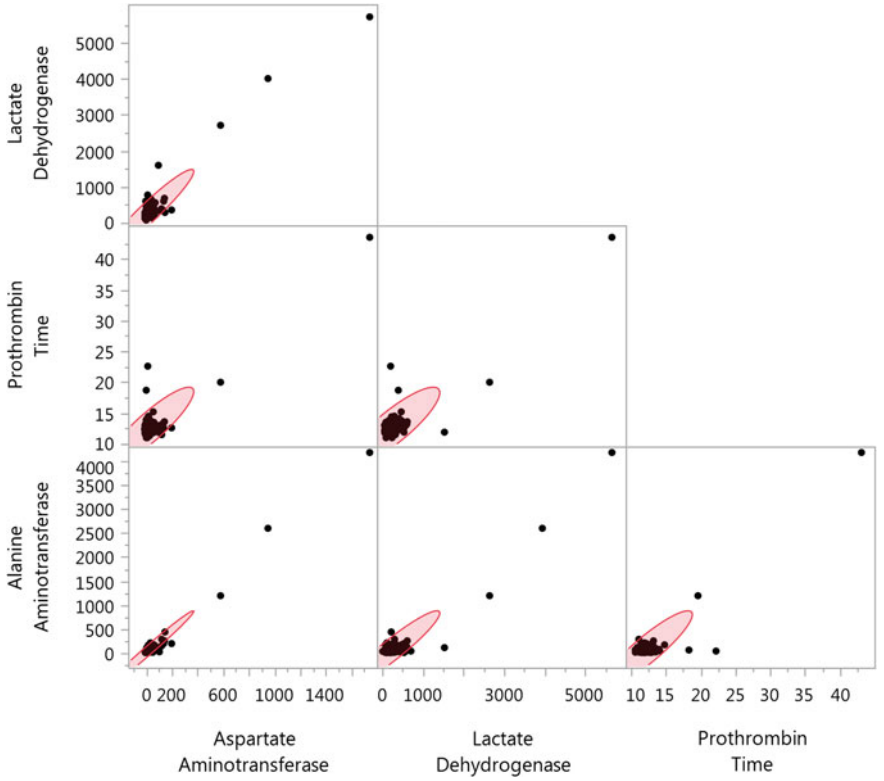


Fig. 13.5 Scatterplot matrix of highly correlated laboratory measurements for site 28. Pink ovals represent the 95% density ellipses of respective pairwise distributions

Table 13.2 Comparison of study day for a particular visit

	18	19	20	21	22	23	24	25
Suspect								
Reference								

visit from the suspect center may stand out, a more definitive understanding is only possible with some investigation by the study team.

Similar analyses to Sect. 13.4.3.1 can be used here to screen study visits for unusual patterns, such as schedules that appear too perfect, or schedules that indicate that a site tends to see patients too early or too late. The latter phenomenon is problematic in that the sites visits will be off schedule, meaning that the measurements obtained are not reflective of the expected drug exposure. For the example in Fig. 13.6, we can define Table 13.2 in order to compute a CMH row mean score statistic.

In general, I sites and J visits will result in as many as $H = I \times J$ comparisons. Assuming 14 visits for Nicardipine results in at most 560 tests. Unfortunately, given

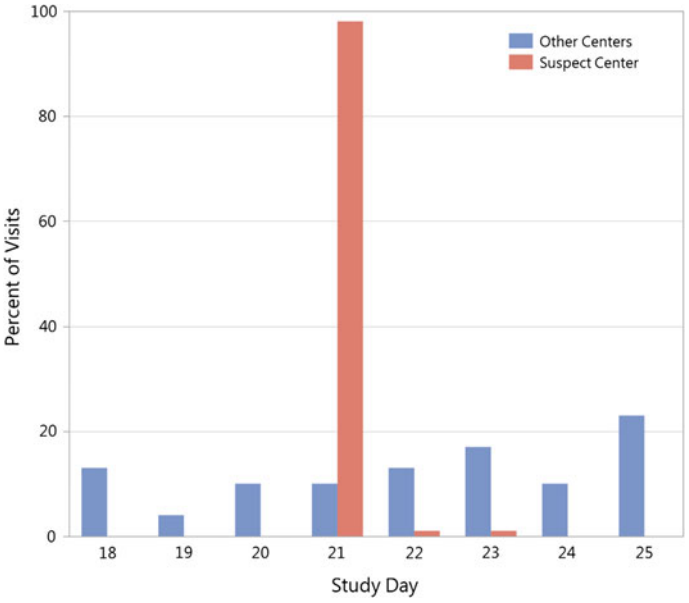


Fig. 13.6 Bar chart of visit attendance for suspect site compared to reference. The y-axis represents the proportion of study visits that occur on the study day summarized along the x-axis. Data are from Buyse et al. 1999.

that each day results in a new visit, the volcano plot, and bar charts for Nicardipine are not extremely interesting. However, this analysis did identify a flaw in my logic in how the SDTM Study Visits (SV) domain data set was built from the vital signs, laboratories, and ECG domains. The laboratories for some patients at Site 39 were performed at Study Day 1 (the day before randomization and dosing), which was subsequently used as the Study Day for Visit 1 and identified as a signal. The importance of this example is that screening for data quality issues generated by the site (or patient) may identify instances where the sponsor, or contract research organization (CRO) working on behalf of the sponsor, may be responsible for the anomaly!

13.5 Conclusions

This chapter summarized past examples of clinical trial misconduct and illustrated several methodologies to assess data quality using data from a sample clinical trial. However, numerous additional methodologies to identify unusual data exist in the literature (e.g., Buyse et al. 1999; Evans 2001; Venet et al. 2012; Kirkwood et al. 2013; TransCelerate BioPharma 2013; Zink et al. 2014; Knepper et al. 2016). Statisticians should take an active interest in data quality, since any steps taken to identify and resolve data anomalies earlier in the timeline of a clinical trial contributes to more

straightforward analysis and interpretation later on. With sufficient resources, study teams should assess the usefulness of various methodologies in the presence of known data issues (O'Kelly 2004). One thing to note from the examples presented above: Identifying unusual data points is only the first step. It takes additional investigation to understand the root cause of any signals indicating a lapse in data quality. While much of the research focuses on the impact of staff at investigative sites, patients, vendors, and sponsors play important roles in data quality. To illustrate this point, we end with a recent example.

In 2013, a scientist at a pharmaceutical services company was convicted of manipulating data for preclinical studies for an anticancer therapy (Anonymous 2013; Benderly 2013; Mansel 2013). The data irregularities were identified in 2009 during the review of quality control analyses within the CRO. It was later discovered that the individual had been engaged in the selective reporting of data since 2003, necessitating the review of hundreds of previously conducted safety studies for multiple sponsors including AstraZeneca and Roche (Jack 2013). This deception "directly impacted the validity of clinical trials and delayed a number of medicines coming to market" (Benderly 2013). It is unclear whether trial sponsors could have identified the data issues, or why it took several years for the CRO to identify the misconduct. However, we repeat the following lesson from the introduction: Defining a series of statistical and graphical checks to be implemented on a regular basis to identify lapses in data quality is a minimal investment to prevent potential catastrophe. Companies should study these examples from the literature and develop safeguards to limit or prevent misconduct and other data quality issues from occurring within and between organizations.

Acknowledgements The author thanks Karl Peace for the invitation to contribute to this volume. This work was conducted while RCZ was an employee of SAS Institute.

References

Akhtar-Danesh, A., & Dehghan-Kooshkghazi, M. (2003). How does correlation structure differ between real and fabricated data-sets? *BMC Medical Research Methodology, 3*(18), 1–9.

Al-Marzouki, S., Evans, S., Marshall, T., & Roberts, I. (2005). Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *British Medical Journal, 331,* 267–270.

Anonymous. (2013, April 17). Scientist Steven Eaton jailed for falsifying drug test results. *BBC News*. http://www.bbc.com/news/uk-scotland-edinburgh-east-fife-22186220.

Baigent, C., Harrell, F. E., Buyse, M., Emberson, J. R., & Altman, D. G. (2008). Ensuring trial validity by data quality assurance and diversification of monitoring methods. *Clinical Trials, 5,* 49–55.

Bailey, K. R. (1991). Detecting fabrication of data in a multicenter collaborative animal study. *Controlled Clinical Trials, 12,* 741–752.

Benderly, B. L. (2013, May 3). A prison sentence for altering data. *Science Careers*. http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2013_05_03/caredit.a1300092.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society, 78,* 551–572.

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

Buyse, M., George, S. L., Evans, S., Geller, N. L., Ranstam, J., Scherrer, B., et al. (1999). The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Statistics in Medicine, 18,* 3435–3451.

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.

Desmet, L., Venet, D., Doffagne, E., Timmermans, C., Burzykowski, T., Legrand, C., et al. (2013). Linear mixed-effects models for central statistical monitoring of multicenter clinical trials. *Statistics in Medicine, 33,* 5265–5279.

Desmet, L., Venet, D., Doffagne, E., Timmermans, C., Legrand, C., Burzykowski, T., et al. (2017). Use of the beta-binomial model for central statistical monitoring of multicenter clinical trials. *Statistics in Biopharmaceutical Research, 9,* 1–11.

European Medicines Agency. (2013). Reflection paper on risk based quality management in clinical trials. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/11/WC500155491.pdf.

Evans, S. (2001). Statistical aspects of the detection of fraud. In S. Lock & F. Wells (Eds.), *Fraud and misconduct in biomedical research*. BMJ Books.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings from the National Academy of Sciences, 109,* 17028–17033.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1,* 3–32.

Fisher, R. A. (1970). *Statistical methods for research workers* (14th ed.). Davien, CT: Hafner Publishing Company.

Grieneisen, M. L., & Zhang, M. (2012). A comprehensive survey of retracted articles from the scholarly literature. *Public Library of Science (PLoS) ONE, 7*(10), 1–15.

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.

Hill, T. P. (1996). A statistical derivation of the significant-digit law. *Statistical Science, 10,* 354–363.

Hsu, J. (1992). Stepwise multiple comparisons with the best. *Journal of Statistical Planning and Inference, 33,* 197–204.

International Conference of Harmonisation. (1996). E6: Guideline for good clinical practice. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6_R1/Step4/E6_R1__Guideline.pdf.

Jack, A. (2013, March 12). Man guilty of manipulating drug tests. *Financial Times*. https://www.ft.com/content/ff8d0e3e-8b25-11e2-8fcf-00144feabdc0#axzz2SAhI5C2S.

Jennett, B., & Bond, M. (1975). Assessment of outcome after severe brain damage: A practical scale. *Lancet, 1,* 480–484.

Kirkwood, A. A., Cox, T., & Hackshaw, A. (2013). Application of methods for central statistical monitoring in clinical trials. *Clinical Trials, 10,* 783–806.

Knepper, D., Lindblad, A. S., Sharma, G., Gensler, G. R., Manukyan, Z., Matthews, A. G., et al. (2016). Statistical monitoring in clinical trials: Best practices for detecting data anomalies suggestive of fabrication or misconduct. *Therapeutic Innovation & Regulatory Science, 50,* 144–154.

Mansel P. (2013, March 14). MHRA successfully prosecutes on preclinical data manipulation. *PharmaTimes*. http://www.pharmatimes.com/news/mhra_successfully_prosecutes_on_preclinical_data_manipulation_1004637?rl=1&rlurl=/13-04-19/UK_pharma_scientist_jailed_over_trial_fraud.aspx.

Office of Research Integrity. (2017). Definition of research misconduct. http://ori.hhs.gov/definition-misconduct.

O'Kelly, M. (2004). Using statistical techniques to detect fraud: A test case. *Pharmaceutical Statistics, 3,* 237–246.

Ott, E. R. (1967). Analysis of means: A graphical procedure. *Industrial Quality Control, 24*, 101–109. Reprinted in 1983 in the *Journal of Quality Technology, 15*, 10–18.

Stokes, M. E., Davis, C. S., & Koch, G. G. (2012). *Categorical data analysis using SAS* (3rd ed.). Cary, NC: SAS Institute Inc.

Taylor, R. N., McEntegart, D. J., & Stillman, E. C. (2002). Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. *Drug Information Journal, 36,* 115–125.

TransCelerate BioPharma Inc. (2013). Position paper: Risk-based monitoring methodology. http://www.transceleratebiopharmainc.com/wp-content/uploads/2016/01/TransCelerate-RBM-Position-Paper-FINAL-30MAY2013.pdf.pdf.

US Food and Drug Administration. (2009). Guidance for industry: Drug-induced liver injury: Premarketing clinical evaluation. https://www.fda.gov/downloads/Drugs/…/guidances/UCM174090.pdf.

US Food & Drug Administration. (2013). Guidance for industry: Oversight of clinical investigations—A risk-based approach to monitoring. Available at: http://www.fda.gov/downloads/Drugs/…/Guidances/UCM269919.pdf.

Venet, D., Doffagne, E., Burzykowski, T., Beckers, F., Tellier, Y., Genevois-Marlin, E., et al. (2012). A statistical approach to central monitoring of data quality in clinical trials. *Clinical Trials, 9,* 705–713.

Weir, C., & Murray, G. (2011). Fraud in clinical trials: Detecting it and preventing it. *Significance, 8,* 164–168.

Westfall, P. H., Tobias, R. D., & Wolfinger, R. D. (2011). *Multiple comparisons and multiple tests using SAS* (2nd ed.). Cary, North Carolina: SAS Institute.

White, C. (2005). Suspected research fraud: Difficulties of getting at the truth. *British Medical Journal, 331,* 281–288.

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. (2014). *Risk-based monitoring and fraud detection in clinical trials using JMP and SAS*. Cary, North Carolina: SAS Institute.

Zink, R. C. (2017). Detecting safety signals among adverse events in clinical trials. In K. E. Peace, D. G. D. Chen, & S. M. Menon (Eds.), *ICSA Biostatistics Book Series of the Biopharmaceutical Applied Statistics Symposium (BASS). Volume 2: Biostatistical Analysis of Clinical Trials*. Springer.

Chapter 14
Design and Analysis of Biosimilar Studies

Shein-Chung Chow and Fuyu Song

14.1 Introduction

In 2009, the United States (US) Congress passed the *Biologics Price Competition and Innovation* (*BPCI*) *Act*, which has given the US Food and Drug Administration (FDA) the authority to review and approve biosimilar drug products (or follow-on biologics). A biosimilar product is a similar biological product such as protein product, vaccine, or blood product whose active drug substance is made of a living cell or derived from a living organism. Biosimilars are not generic drugs but *similar* biologic drug products. Similar is in the sense that it is similar to an innovator drug product in terms of safety, purity, and potency.

Following the passage of the BPCI Act, in order to obtain input on specific issues and challenges associated with the implementation of the BPCI Act, the US FDA conducted a two-day public hearing on *Approval Pathway for Biosimilar and Interchangeability Biological Products* held on November 2–3, 2010 at the FDA in Silver Spring, Maryland. Several scientific factors were raised and discussed at the public hearing. These scientific factors include criteria for assessing biosimilarity, study design and analysis methods for assessment of biosimilarity, and tests for comparability in quality attributes of manufacturing process and/or immunogenicity (see, e.g., Chow et al. 2010). These issues primarily focus on the assessment of biosimilarity. The issue of interchangeability in terms of the concepts of alternating and switching was also mentioned and discussed. The discussions of these scientific factors have led to the development of regulatory guidances. On February 9, 2012, the US FDA

S.-C. Chow
Duke University School of Medicine, Durham, NC, USA

F. Song (✉)
Center for Food and Drug Inspection, China Food and Drug Administration, No. 11 Building, Fahua Nanli, Dongcheng, Beijing 100161, China
e-mail: songfy@cfdi.org.cn

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_14

circulated three draft guidances on the demonstration of biosimilarity for comments. These draft guidances are

(1) Scientific Considerations in Demonstrating Biosimilarity to a Reference Product (FDA 2012a);
(2) Quality Considerations in Demonstrating Biosimilarity to a Reference Protein Product (FDA 2012b);
(3) Biosimilars: Questions and Answers Regarding Implementation of the Biologics Price Competition and Innovation (BPCI) Act of 2009 (FDA 2012c).

Subsequently, FDA hosted another public hearing on the discussion of these draft guidances at the FDA on May 11, 2012.

As indicated in the guidance of *Scientific Considerations in Demonstrating Biosimilarity to a Reference Product*, FDA recommends a stepwise approach for providing so-called totality-of-the-evidence in demonstrating biosimilarity between a proposed biosimilar to a reference product. The stepwise approach starts with analytical similarity assessment for critical quality attributes (CQAs) at various stages of the manufacturing process, followed by animal studies for toxicity, pharmacokinetic/pharmacodynamics (PK/PD), immunogenicity, and clinical studies for safety and efficacy assessment. The purpose of this chapter is to provide a comprehensive review of design and analysis for biosimilar studies conducted for biosimilarity assessment, analytical similarity assessment, and assessment of the risk of switching and/or alternation for drug interchangeability.

In the next section, criteria and commonly considered statistical methods (Schuirmann's two one-sided tests procedure and confidence interval approach) for biosimilarity assessment will be reviewed. Also, included in this section is a biosimilarity index proposed by Chow et al. (2011). Section 14.3 focuses on tiered approach recommended by the FDA for analytical similarity assessment. Issues and study designs regarding drug interchangeability in terms of switching and alternation are discussed in Sect. 14.4. Recent development and some concluding remarks are provided in the last section of this chapter.

14.2 Assessing Biosimilarity of Biosimilar Products

14.2.1 Definition of Biosimilarity

As indicated in the BPCI Act, a biosimilar product is a product that is *highly similar* to the reference product notwithstanding *minor* differences in clinically inactive components and there are no clinically meaningful differences in terms of safety, purity, and potency. At 2010 FDA public hearing, the following scientific issues regarding design and analysis of biosimilar studies were raised.

(1) *How similar is similar?* and *How similar is considered highly similar?*

Table 14.1 Fundamental differences between generics and biosimilars

Generic drug products	Biosimilar drug products
Made by chemical synthesis	Made by living cells
Defined structure	Heterogeneous structure Mixtures of related molecules
Easy to characterize	Difficult to characterize
Relatively stable	Variable Sensitive to environmental conditions such as light and temperature
No issue of immunogenicity	Issue of immunogenicity
Usually taken orally	Usually injected
Often prescribed by a general practitioner	Usually prescribed by specialists

(2) Will *one-size-fits-all* criterion for bioequivalence assessment of generic drug products be adopted for biosimilarity assessment of biosimilar products? If not, what is the *ideal* criterion for biosimilarity?
(3) Can *scaled average bioequivalence* (SABE) criterion for highly variable drug products be used for biosimilar products?
(4) Can a *non-inferiority trial* be considered to replace a bioequivalence (biosimilarity) study?
(5) Should a *crossover design* or a *parallel design* be used for biosimilarity assessment?
(6) Can the standard methods for bioequivalence assessment be directly applied to assess biosimilarity?

These questions, however, were not fully addressed at the public hearing, and some of the questions still remain unanswered up to date. In what follows, some of these scientific factors will be addressed.

14.2.2 Fundamental Differences Between Generics and Biosimilars

Since standard methods for bioequivalence assessment of generic drug products are well established and have been in practice for years, the questions regarding whether these methods can be directly applied to assessment of biosimilarity of biosimilar products. The answer to this question is obvious due to some fundamental differences between generics and biosimilars (see Table 14.1).

To provide a better understanding, Table 14.2 summarizes the comparison between in vivo bioequivalence testing for generic drug products and biosimilarity testing for biosimilar products.

Table 14.2 Comparison between in vivo bioequivalence testing and biosimilarity testing

Characteristics	Bioequivalence testing	Biosimilarity testing
Study endpoint	Drug absorption	Drug safety/purity/potency
Variability	20–30%	40–50%
Criterion	(80, 125%)	(70, 143%) or SABE?
Study design	Crossover	Parallel/crossover
Statistical methods	TOST or confidence interval	TOST, confidence interval or biosimilarity index
Primary assumption	Fundamental BE assumption	Fundamental BS assumption?
Requirement	BE trial is required	Analytical, PK/PD, clinical studies

14.2.3 Criteria for Biosimilarity

Average Bioequivalence (ABE) Criterion—For approval of generic drug products, FDA requires that evidence of equivalence I average bioavailability in terms of drug absorption be provided through the conduct of bioequivalence studies. A test drug product is said to be bioequivalent to a reference drug product if the estimated 90% confidence interval for the geometric means ratio (GMR) of the primary pharmacokinetic (PK) parameters, e.g., area under the blood or plasma concentration-time curve (AUC) and maximum concentration (Cmax) is totally within the bioequivalence limits of 80.00–125.00% (FDA 2003; Chow and Liu 2008). ABE criterion focuses on average bioavailability and ignores the variability associated with the PK responses. Thus, two drug products may fail the evaluation of ABE if the variability associated with the PK responses is large even though they have identical means.

ABE criterion has been criticized penalizing good products (i.e., test products with less variability). In this case, there is a need for alternative criteria for bioequivalence assessment for drug products with large variability. A drug with large variability is considered highly variable. FDA defines a highly variable drug (HVD) as a drug whose within-subject (or intra-subject) variation is greater than or equal to 30%. Based on this definition, most biosimilar products are considered highly variable drug products. One of problematic aspects of this definition is that the estimated within-subject variability depends on the metrics of pharmacokinetic responses such as AUC and Cmax. Haidar et al. (2008) pointed out that HVDs show variable pharmacokinetics as a result of their inherent properties (e.g., distribution, systemic metabolism, and elimination) (see also, Haidar et al. 2008; Tothfalusi et al. 2009; Davit et al. 2008). A drug may have low variability if it is administered intravenously, whereas it can be highly variable after oral administration.

Scaled Average Bioequivalence (SABE) Criterion—In practice, HVDs often fail to meet current regulatory acceptance criteria for ABE. In the past decade, the topic for evaluation of bioequivalence for HVDs has received much attention. This topic has been discussed several times at regulatory forums and international conferences,

but academics, representatives of pharmaceutical industries and regulatory agencies failed to reach a consensus until recently that the approach of scaled average bioequivalence (SABE) is proposed by Haidar et al. (2008). The approach of SABE is briefly described below.

Denoted by μ_T and μ_R respectively, are compared. The acceptance of bioequivalence is claimed it the difference between the logarithmic means is between prespecified regulatory limits. The limits (δ_A) are generally symmetrical on the logarithmic scale and usually equal $\pm \ln(1.25)$. Thus, the criterion for ABE can be expressed as follows:

$$-\delta_A \leq \mu_T - \mu_R \leq \delta_A$$

In a bioequivalence study, the individual kinetic responses are evaluated from the measured concentrations. The means of the logarithmic responses of the two formulations are calculated. These sample averages estimate the true population means. A variance is also estimated for each kinetic response. It is a measure of the intra-subject variance but not always identical to it. FDA suggests the above ABE could be scaled by a standard deviation as follows:

$$-\delta_s \leq \frac{(\mu_T - \mu_R)}{\sigma_W} \leq \delta_s \ ,$$

where δ_S is the SABE regulatory cutoff. Here, the standard deviation (σ_W) is the within-subject standard deviation. In replicate design, σ_W is generally the within-subject standard deviation of the reference formulation (denoted by σ_{WR}).

14.2.4 Statistical Methods

In practice, one of the most widely used designs for assessing biosimilarity between biosimilar products and an innovator biological product is probably either a two-sequence, two-period (2×2) crossover design or a two-arm parallel group design. Under a valid study design, biosimilarity can then be assessed by means of an equivalence test under the following interval hypotheses

$$H_0 : \mu_T - \mu_R \leq \theta_L or \mu_T - \mu_R \geq \theta_U \quad \text{vs.} \quad H_a : \theta_L < \mu_T - \mu_R < \theta_U, \qquad (14.1)$$

where (θ_L, θ_U) are pre-specified equivalence limits (margins) and μ_T and μ_R are the population means of a biological (test) product and an innovator biological (reference) product, respectively. Under a crossover design, consider the following statistical model for raw data:

$$Y_{ijk} = \mu + S_{ik} + P_j + F_{j,k} + C_{(j-1,k)} + e_{ijk}, \qquad (14.2)$$

where

Y_{ijk} be the response (e.g., AUC) of the ith subject in the kth sequence at the jth period and

μ the overall mean;

S_{ik} the random effect of the ith subject in the kth sequence, where $i = 1, 2, \ldots, g$;

P_j the fixed effect of the jth period, where $j = 1, \ldots,$ p and $\Sigma_j P_j = 0$;

$F_{(j,\,k)}$ the direct fixed effect of the drug product in the kth sequence which is administered at the jth period, and $\Sigma F_{(j,\,k)} = 0$;

$C_{(j-1,\,k)}$ the fixed first-order carryover effect of the drug product in the kth sequence which is administered at the $(j = 1)$th period, where $C_{(0,\,k)} = 0$; and $\Sigma C_{(j-1,\,k)} = 0$;

e_{ijk} the (within-subject) random error in observing Y_{ijk}.

It is assumed that $\{S_{ik}\}$ are independently and identically distributed (i.i.d.) with mean 0 and variance σ_s^2 and $\{e_{ijk}\}$ are independently distributed with mean 0 and variances σ_t^2, where $t = 1, 2, \ldots, L$ (the number of formulations to be compared). $\{S_{ik}\}$ and $\{e_{ijk}\}$ are assumed mutually independent. The estimate of σ_s^2 is usually used to explain the inter-subject variability, and the estimates of σ_t^2 are used to assess the intra-subject variabilities for the t-th drug product.

Confidence Interval Approach—For bioequivalence assessment of small molecule drug products, the FDA adopts the 80/125 rule based on log-transformed data. The 80/125 rule states that bioequivalence is concluded if the geometric means ratio (GMR) between the test product and the reference product is within the bioequivalence limits of (80.00, 125.00%), with a certain statistical assurance. Thus, a typical approach is to consider the method of classic (shortest) confidence interval.

Consider a standard two-sequence, two-period crossover design, under model (14.2), after log transformation of the data, let $\bar{Y}_T$ and $\bar{Y}_R$ be the respective least squares means for the test and reference formulations, which can be obtained from the sequence-by-period means. The classic (or shortest) $(1 - 2\alpha) \times 100\%$ confidence interval can then be obtained based on the following t statistic:

$$T = \frac{(\bar{Y}_T - \bar{Y}_R) - (\mu_T - \mu_R)}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{14.3}$$

where n_1 and n_2 are the numbers of subjects in sequences 1 and 2, respectively, and $\hat{\sigma}_d$ is an estimate of the variance of the period differences for each subject within each sequence, which are defined as follows:

$$d_{ik} = \frac{1}{2}(Y_{i2k} - Y_{i1k}), \qquad i = 1, 2, \ldots, n_k; \quad k = 1, 2.$$

Thus, $V(d_{ik}) = \sigma_d^2 = \sigma_e^2/2$. Under normality assumptions, T follows a central student t distribution with degrees of freedom $n_1 + n_2 - 2$. Thus, the classic $(1 - 2\alpha) \times 100\%$ confidence interval for $\mu_T - \mu_R$ can be obtained as follows:

$$L_1 = (\bar{Y}_T - \bar{Y}_R) - t(\alpha, n_1 + n_2 - 2)\hat{\sigma}_d\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$
$$U_1 = (\bar{Y}_T - \bar{Y}_R) - t(\alpha, n_1 + n_2 - 2)\hat{\sigma}_d\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{14.4}$$

The above a $(1 - 2\alpha) \times 100\%$ confidence interval for $\log(\mu_T) - \log(\mu_R) = \log(\mu_T/\mu_R)$ can be converted into a $(1 - 2\alpha) \times 100\%$ confidence interval for μ_T/μ_R by taking an anti-log transformation.

Note that under a parallel group design, a $(1 - 2\alpha) \times 100\%$ confidence interval for μ_T/μ_R can be similarly obtained.

Schuirmann's Two One-sided Tests (TOST) Procedure—The assessment of average bioequivalence is based on the comparison of bioavailability profiles between product products. However, in practice, it is recognized that no two drug products will have exactly the same bioavailability profiles. Therefore, if the profiles of the two drug products differ by less than a (clinically) meaningful limit, the profiles of the two drug products may be considered equivalent. Following this concept, Schuirmann (1981) first introduced the use of interval hypotheses (14.1) for assessing average bioequivalence. The concept of interval hypotheses (14.1) is to show average bioequivalence by rejecting the null hypothesis of average bio*in*equivalence. In most bioavailability and bioequivalence studies, δ_L and δ_U are often chosen to be $-\theta_L = \theta_U = 20\%$ of the reference mean (μ_R). When the natural logarithmic transformation of the data is considered, the hypotheses corresponding to hypotheses (14.1) can be stated as

$$\begin{aligned} &H_0': \mu_T/\mu_R \le \delta_L \quad \text{or} \quad \mu_T/\mu_R \ge \delta_U \\ \text{vs. } &H_a': \delta_L < \mu_T/\mu_R < \delta_U \end{aligned} \tag{14.5}$$

where $\delta_L = \exp(\theta_L)$ and $\delta_U = \exp(\theta_U)$. Note that FDA recommends that $(\delta_L, \delta_U) = (80.00, 125.00\%)$ for assessing average bioequivalence.

Note that the test for hypotheses in (14.5) formulated on the log-scale is equivalent to testing for hypotheses (14.1) on the raw scale. The interval hypotheses (14.1) can be decomposed into two sets of one-sided hypotheses

$$H_{01}: \mu_T - \mu_R \le \theta_L \quad \text{vs.} \quad H_{a1}: \mu_T - \mu_R > \theta_L$$

and

$$H_{02}: \mu_T - \mu_R \ge \theta_U \quad \text{vs.} \quad H_{a2}: \mu_T - \mu_R < \theta_U. \tag{14.6}$$

The first set of hypotheses is to verify that the average bioavailability of the test formulation is not too low, whereas the second set of hypotheses is to verify that the average bioavailability of the test formulation is not too high. A relatively low (or high) average bioavailability may refer to the concern of efficacy (or safety) of the test formulation. If one concludes that $\theta_L < \mu_T - \mu_R$ (i.e., reject H_{01}) and $\mu_T - \mu_R < \theta_U$ (i.e., reject H_{02}), then it has been concluded that

$$\theta_L < \mu_T - \mu_R < \theta_U.$$

Thus, μ_T and μ_R are equivalent. The rejection of H_{01} and H_{02}, which leads to the conclusion of average bioequivalence, is equivalent to rejecting H_0 in (14.1).

Under hypotheses (14.1), Schuirmann (1987) introduced the two one-sided tests procedure for assessing average bioequivalence between drug products. The proposed two one-sided tests procedure suggests the conclusion of equivalence of μ_T and μ_R at the α level of significance if, and only if, H_{01} and H_{02} in (14.6) are rejected at a pre-determined α-level of significance. Under the normally assumptions, the two sets of one-sided hypotheses can be tested with ordinary one-sided t tests. We conclude that of μ_T and μ_R are average equivalent if

$$T_L = \frac{(\bar{Y}_T - \bar{Y}_R) - \theta_L}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t(\alpha, n_1 + n_2 - 2)$$

and

$$T_U = \frac{(\bar{Y}_T - \bar{Y}_R) - \theta_U}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t(\alpha, n_1 + n_2 - 2). \tag{14.7}$$

The two one-sided t tests procedure is operationally equivalent to the classic (shortest) confidence interval approach; that is, both the classic confidence interval approach and Schuirmann's two one-sided tests procedure will lead to the same conclusion on bioequivalence.

Note that under a parallel group design, Schuirmann's two one-sided tests procedure can be similarly derived with a slightly modification from a pair t test statistic to a two-sample t test statistic.

14.2.5 Biosimilarity Index

Chow (2013) proposed the development of a composite index for assessing the biosimilarity of follow-on biologics based on the facts that (1) the concept of biosimilarity for biologic products (made of living cells) is very different from that of bioequivalence for drug products, and (2) biologic products are very sensitive to small changes in the variation during the manufacturing process (i.e., it might have

a drastic change in clinical outcome). Although some research on the comparison of moment-based criteria and probability-based criteria for the assessment of (1) average biosimilarity and (2) variability of biosimilarity for some given study endpoints by applying the criteria for bioequivalence are available in the literature (see, e.g., Chow et al. 2010; Hsieh et al. 2010), universally acceptable criteria for biosimilarity are not available in the regulatory guidelines/guidances. Thus, Chow (2013) and Chow et al. (2011) proposed a biosimilarity index based on the concept of the probability of reproducibility as follows.

Step 1. Assess the average biosimilarity between the test product and the reference product based on a given biosimilarity criterion. For illustration purpose, consider bioequivalence criterion as biosimilarity criterion. That is, biosimilarity is claimed if the 90% confidence interval of the ratio of means of a given study endpoint falls within the biosimilarity limit of (80.00, 125.00%) or (−0.2231, 0.2231) based on log-transformed data or based on raw (original) data.

Step 2. Once the product passes the test for biosimilarity in Step 1, calculate the reproducibility probability based on the observed ratio (or observed mean difference) and variability. Thus, the calculated reproducibility probability will take the variability and the sensitivity of heterogeneity in variances into consideration for assessment of biosimilarity.

Step 3. We then claim biosimilarity if the calculated 95% confidence lower bound of the reproducibility probability is larger than a pre-specified number p_0, which can be obtained based on an estimated of reproducibility probability for a study comparing a "reference product" to itself (the "reference product"). We will refer to such a study as an R-R study. Alternatively, we can then claim (local) biosimilarity if the 95% confidence lower bound of the biosimilarity index is larger than p_0.

Since the idea of the biosimilar index is to show that the reproducibility probability in a study for comparing "a reference product" with "the reference product" is higher than the study for comparing a follow-on biologic with the innovative (reference) product, the criterion of an acceptable reproducibility probability (i.e., p_0) for assessment of biosimilarity can be obtained based on the R-R study. For example, if the R-R study suggests the reproducibility probability of 90%, i.e., $P_{RR} = 90\%$, the criterion of the reproducibility probability for bioequivalence study could be chosen as 80% of the 90% which is $p_0 = 80\% \times P_{RR} = 72\%$.

The above described biosimilar index has the advantages that (1) it is robust with respect to the selected study endpoint, biosimilarity criteria, and study design, and (2) the probability of reproducibility will reflect the sensitivity of heterogeneity in variance.

Note that the proposed biosimilarity index can be applied to different functional areas (domains) of biological products such as pharmacokinetics (PK), biological activities, biomarkers (e.g., pharmacodynamics), immunogenicity, manufacturing process, efficacy, etc. An overall biosimilarity index or totality biosimilarity index across domains can be similarly obtained as follows:

Step 1. Obtain $\hat{p}_i$, the probability of reproducibility for the ith domain, $i = 1, \ldots, K$.
Step 2. Define the biosimilarity index $\hat{p} = \sum_{i=1}^{K} w_i \hat{p}_i$, where w_i is the weight for the ith domain.
Step 3. Claim global biosimilarity if we reject the null hypothesis that $p \leq p_0$, where p_0 is a pre-specified acceptable reproducibility probability. Alternatively, we can claim (global) biosimilarity if the 95% confidence lower bound of p is larger than p_0.

14.3 Analytical Similarity Assessment

In the guidance on *Scientific Considerations in Demonstrating Biosimilarity to a Reference Product*, FDA introduces the concept of stepwise approach for obtaining totality-of-the-evidence for the regulatory review and approval of biosimilar applications (FDA 2015).

The stepwise approach starts with the assessment of analytical similarity of critical quality attributes (CQAs) for structural and functional characterization in the manufacturing process of biosimilar product that may have an impact on assessment of similarity. In practice, there is often a large number of CQAs that may be relevant to clinical outcomes. Thus, it is almost impossible to assess analytical similarity for all of these CQAs individually. As a result, FDA suggests that the sponsors identify CQAs that are relevant to clinical outcomes and classify them into three tiers depending upon their criticality risk ranking, i.e., most relevant (Tier 1), mild-to-moderately relevant (Tier 2), and least relevant (Tier 3) to clinical outcomes. To assist the sponsors, FDA also proposes some statistical approaches for the assessment of analytical similarity for CQAs from different tiers. For example, FDA recommends equivalence test for CQAs from Tier 1, a quality range approach for CQAs from Tier 2, and descriptive raw data and graphical presentation for CQAs from Tier 3 (see, e.g., Christl 2015; Tsong 2015; Chow 2014, 2015).

14.3.1 Stepwise Approach for Demonstrating Biosimilarity

The stepwise approach is briefly summarized by a pyramid illustrated in Fig. 14.1.

The stepwise approach starts with analytical studies for structural and functional characterization. The stepwise approach continues with animal studies for toxicity, clinical pharmacology studies such as PK/PD studies, followed by investigations of immunogenicity, and clinical studies for safety/tolerability and efficacy. The sponsors are encouraged to consult with medical/statistical reviewers of FDA with the proposed plan or strategy of the stepwise approach for regulatory agreement and acceptance. This is to make sure that the information provided is sufficient to fulfill the FDA's requirement for providing totality-of-the-evidence for the demonstration of biosimilarity of the proposed biosimilar product as compared to the reference prod-

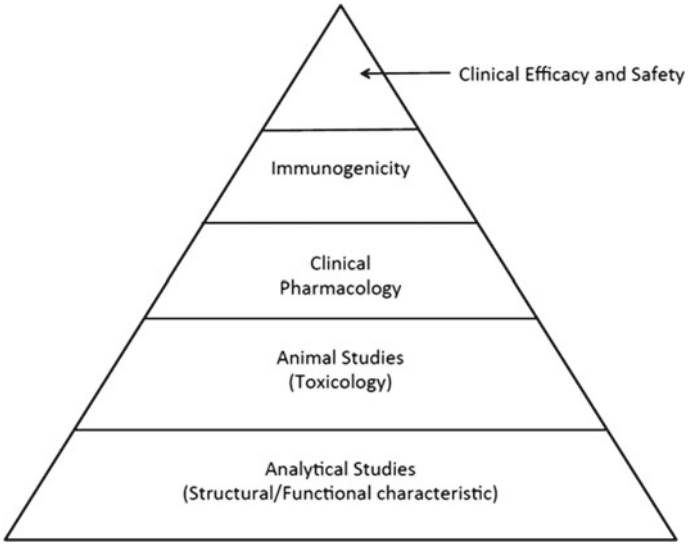


Fig. 14.1 A stepwise approach to demonstrate biosimilarity

uct. FDA suggests critical quality attributes be classified into three tiers depending upon their criticality or risk ranking relevant to the clinical outcomes.

Tsong (2015) indicated that critical quality attributes (CQAs) are necessarily tested for the functional, structural, and physicochemical characterization of the proposed biosimilar product as compared to a reference product (either a US-licensed product or an EU-approved reference product) for analytical similarity assessment. Analytical similarity assessment is considered as the foundation of the stepwise approach for obtaining the totality-of-the-evidence for demonstrating biosimilarity between the proposed biosimilar product and the reference product. The CQAs with most relevance to clinical outcomes will be assigned to Tier 1, while the CQAs with mild-to-moderately relevant to clinical outcomes will be classified to Tier 2. Tier 3 will contain those CQAs with least relevance to clinical outcomes. In practice, it is believed that biological activity assays are the best representation available to test the clinically relevant mechanism of action (MOA) and therefore should be assigned to Tier 1. Other CQAs which are tested in comparative physicochemical and functional assessment (outside of those relevant to MOA) are of potential relevance to similarity which is considered most appropriate for Tier 2 or Tier 3.

14.3.2 FDA's Tiered Approach

FDA recommends tiered approach for analytical similarity assessment of CQAs from different tiers. For CQAs from Tier 1, FDA recommends an equivalence test be

performed. For CQAs from Tier 2, it is suggested that quality range approach should be considered. For CQAs from Tier 3, descriptive raw data and graphical presentation be used (see, e.g., Cristl 2015; Tsong 2015; Chow 2015). These approaches for tier analysis are briefly outlined in below.

Equivalence Test for Tier 1—For CQAs from Tier 1, FDA recommends that an equivalency test be performed for the assessment of analytical similarity. As indicated by the FDA, a potential approach could be a similar approach to bioequivalence testing for generic drug products (FDA 2003; Chow 2015). In other words, for a given critical attribute, we may test for equivalence by the following interval (null) hypothesis:

$$H_0 : \mu_T - \mu_R \leq -\delta \quad \text{or} \quad \mu_T - \mu_R \geq \delta, \tag{14.8}$$

where $\delta > 0$ is the equivalence limit (or similarity margin), and μ_T and μ_R are the mean responses of the test (the proposed biosimilar) product and the reference product lots, respectively. Analytical equivalence (similarity) is concluded if the null hypothesis of non-equivalence (*dis*-similarity) is rejected. Note that Yu (2004) defined inequivalence as when the confidence interval falls entirely outside the equivalence limits. Similarly to the confidence interval approach for bioequivalence testing under the raw data model, analytical similarity would be accepted for a quality attribute if the $(1 - 2\alpha)$ 100% two-sided confidence interval of the mean difference is within $(-\delta, \delta)$.

Under the null hypothesis (14.8), FDA indicates that the equivalence limit (similarity margin), δ, would be a function of the variability of the reference product, denoted by σ_R. It should be noted that each lot contributes one test value for each attribute being assessed. Thus, σ_R is the population standard deviation of the lot values of the reference product. Ideally, the reference variability, σ_R, should be estimated based on some sampled lots randomly selected from a pool of reference lots for the statistical equivalence test. In practice, it may be a challenge when there is a limited number of available lots. Thus, FDA suggests the that the sponsor provide a plan on how the reference variability, σ_R, will be estimated with a justification.

Quality Range Approach for Tier 2—For CQAs from Tier 2, FDA suggests that analytical similarity be assessed on the basis of the concept of quality ranges, i.e., $\pm x\sigma$, where σ is the standard deviation of the reference product and x should be appropriately justified. Thus, the quality range of the reference product for a specific quality attribute is defined as $\left(\hat{\mu}_R - x\hat{\sigma}_R, \hat{\mu}_R + x\hat{\sigma}_R\right)$. Analytical similarity would be accepted for the quality attribute if a sufficient percentage of test lot values (e.g., 90%) falls within the quality range. For a given critical attribute the quality range is set based on test results of available reference lots. If $x = 1.645$, we would expect 90% of the test results from reference lots to lie within the quality range. If x is chosen to be 1.96, we would expect that about 95% test results of reference lots will fall within the quality range. As a result, the selection of x could impact the quality range and consequently the percentage of test lot values that will fall within the

quality range. Thus, FDA indicates that the standard deviation multiplier (x) should be appropriately justified.

The quality range approach for comparing populations between a proposed biosimilar product and a reference product is a reasonable approach under the assumption that $\mu_T = \mu_R$ and $\sigma_T = \sigma_R$. Under this assumption, we expect that there is a high percentage (say 90%) of test values of the test product will fall within the quality range obtained based on the test values of the reference product. Thus, one of the major criticisms of the quality range approach is that it ignores the fact that there are differences in population mean and population standard deviation between the proposed biosimilar product and the reference product, i.e., $\mu_T \neq \mu_R$ and $\sigma_T \neq \sigma_R$. In practice, it is recognized that biosimilarity between a proposed biosimilar product and a reference product could be established even under the assumption that $\mu_T \neq \mu_R$ and $\sigma_T \neq \sigma_R$. Thus, under the assumption that $\mu_T = \mu_R$ and $\sigma_T = \sigma_R$, the quality range approach for analytical similarity assessment for CQAs from Tier 2 is considered more stringent as compared to equivalence testing for CQAs from Tier 1 (most relevant to clinical outcomes) regardless that they are mild-to-moderately relevant to clinical outcomes. This is because that equivalence testing allows a possible mean shift of $\sigma_R/8$, while the quality range approach does not. In what follows, several examples for the possible scenarios of (1 $\mu_T \approx \mu_R$ or there is a significant mean shift (either a shift to the right or a shift to the left), and (2) $\sigma_T \approx \sigma_R$, $\sigma_T > \sigma_R$, or $\sigma_T < \sigma_R$.

Raw data and graphical comparison for Tier 3—For CQAs in Tier 3 with lowest risk ranking, FDA recommends an approach that uses raw data/graphical comparisons. The examination of similarity for CQAs in Tier 3 is by no means less stringent, which is acceptable because they have least impact on clinical outcomes in the sense that a notable dissimilarity will not affect clinical outcomes.

The method of raw data and graphical comparison is easy to implement and yet it is subjective. One of the major criticisms is that it is not clear how the approach can provide totality-of-the-evidence for demonstrating biosimilarity. For CQAs in Tier 1, they are least relevant to clinical outcomes and yet should carry less weight as compared to those CQAs from Tier 1 and Tier 2. There is little or no information regarding what results will be accepted by the method of data and graphical comparison. In practice, if significant differences in graphical comparisons of some CQAs are observed, should this observation raise a concern? In this case, if it is possible, the degree of criticality risk ranking of these CQAs should be assessed whenever possible.

14.3.3 Challenging Issues to FDA's Approaches

The idea of FDA's proposed equivalence test for Tier 1 CQAs comes from the bioequivalence assessment for generic drugs which contain the same active ingredient(s) as the reference drug product. It may not be appropriate to apply the idea directly to

the assessment of biosimilarity of biosimilar products. The FDA's proposed equivalence test is sensitive to (1) the primary assumptions made, (2) the selection of c, and (3) the estimation of σ_R. Chow (2015) commented on these issues as follows.

Primary Assumptions—Basically, FDA's proposed equivalence test ignores (1) the lot-to-lot variability of both the reference product and the proposed biosimilar product, (2) the difference between means, and (3) the inflation/deflation in variability between the reference product and the proposed biosimilar product. Suppose that there are K reference lots which will be used to establish EAC for the equivalence test. FDA suggests that one sample is randomly selected from each lot. The standard deviation of the reference product σ_R can be estimated based on the K test results. Let $x_i, i = 1, 2., \ldots, K$ be the test result of the ith lot. $x_i, i = 1, 2, \ldots, K$ are assumed to be independently and identically distributed with mean μ_R and variance σ_R^2. In other words, we assume that $\mu_{Ri} = \mu_{Rj} = \mu_R$ and $\sigma_{Ri}^2 = \sigma_{Rj}^2 = \sigma_R^2$ for $i \neq j, i, j = 1, 2, \ldots, K$. Thus, the expected value of $E(\bar{x}) = \mu_R$ and $Var(\bar{x}) = \sigma_R^2/K$. In practice, it is well recognized that $\mu_{Ri} \neq \mu_{Rj}$ and $\sigma_{Ri}^2 \neq \sigma_{Rj}^2$ for $i \neq j$, where μ_{Ri} and σ_{Ri}^2 are the mean and variance of the ith lot of the reference product. A similar argument applies to the proposed biosimilar (test) product. As a result, the selection of reference lots for the estimation of σ_R is critical for the proposed approach.

In addition, FDA assumes that the difference in mean responses between the reference product and the proposed biosimilar product is proportional to the variability of the reference product. In other words, $\Delta = \mu_T - \mu_R$(in log scale) $\propto \sigma_R$. FDA suggests that the power for detecting a clinically meaningful difference be evaluated at $\sigma_R/8$. Thus, under the assumption, the FDA's proposed equivalence testing is straightforward and easy to implement. However, Chow (2014) indicated that FDA's proposed testing procedure depends upon the selection of the regulatory standard $c = 1.5$, the anticipated difference $\Delta = \mu_T - \mu_R$, and the compromise between the test size (type I error) and statistical power (type II error) for detecting Δ Chow 2015).

Heterogeneity Within and Between Test and Reference Products—Let σ_R^2 and σ_T^2 be the variabilities associated with the reference product and the test product, respectively. Also, let n_R and n_T be the number of lots for analytical similarity assessment for the reference product and the test product, respectively. Thus, we have

$$\sigma_R^2 = \sigma_{WR}^2 + \sigma_{BR}^2 \quad \text{and} \quad \sigma_T^2 = \sigma_{WT}^2 + \sigma_{BT}^2,$$

where $\sigma_{WR}^2, \sigma_{BR}^2$ and $\sigma_{WT}^2, \sigma_{BT}^2$ are the within-lot variability and between-lot (lot-to-lot) variability for the reference product and the test product, respectively. In practice, it is very likely that $\sigma_R^2 \neq \sigma_T^2$ and often $\sigma_{WR}^2 \neq \sigma_{WT}^2$ and $\sigma_{BR}^2 \neq \sigma_{BT}^2$ even $\sigma_R^2 \approx \sigma_T^2$. This has posted a major challenge to the FDA's proposed approaches for the assessment of analytical similarity for CQAs from both Tier 1 and Tier, especially when there is only one test sample from each lot from the reference product and the test product. FDA's proposal ignores lot-to-lot (between-lot) variability, i.e., when

$\sigma_{BR}^2 = 0$ or $\sigma_R^2 = \sigma_{WR}^2$. In other words, sample variance based on $x_i, i = 1, \ldots, K$ from the reference product may underestimate the true σ_R^2, and consequently may not provide a fair and reliable assessment of analytical similarity for a given quality attribute.

In practice, it is well recognized that $\mu_{Ri} \neq \mu_{Rj}$ and $\sigma_{Ri}^2 \neq \sigma_{Rj}^2$ for $i \neq j$, where μ_{Ri} and σ_{Ri}^2 are the mean and variance of the ith lot of the reference product. A similar argument is applied to the proposed biosimilar (test) product. As a result, the selection of reference lots for the estimation of σ_R is critical for the proposed approach. The selection of reference lots has an impact on the estimation of σ_R and consequently on the EAC. Suppose there are K reference lots available and n lots will be tested for analytical similarity. FDA suggests using the remaining K-n lots to establish EAC to avoid selection bias. It sounds a reasonable approach if $K \gg n$. In practice, however, there are few lots available. In this case, the FDA's proposed approach may not be feasible.

Sample Size—In practice, one of the major problems to a biosimilar sponsor is the availability of reference lots for analytical similarity testing. FDA suggests that an appropriate sample size (the number of lots from the reference product and from the test product) be used for achieving a desired power (say 80%) to establish similarity based on a two-sided test at the 5% level of significance assuming that the mean response of the test product differs from that of the reference product by $\sigma_R/8$.

Furthermore, since sample size is a function of α (type I error), β (type II error or 1 minus power), δ (treatment effect), and σ^2 (variability), it is a concern that we may have inflated the type I error rate for achieving a desired power to detect a clinically meaningful effect size (adjusted for variability) with a pre-selected small sample size (i.e., a small number of lots).

14.4 Issues and Designs of Drug Interchangeability

As indicated in the Subsection (b)(3) amended to the Public Health Act Subsection 351(k)(3), the term *interchangeable* or *interchangeability* in reference to a biological product that is shown to meet the standards described in Subsection (k)(4), means that the biological product may be substituted for the reference product without the intervention of the healthcare provider who prescribed the reference product. Along this line, in what follows, definition and basic concepts of interchangeability (in terms of switching and alternating) are given.

14.4.1 Definition and Basic Concepts

As indicated in the Subsection (a)(2) amends the Public Health Act Subsection 351(k)(3), a biological product is considered to be interchangeable with the

reference product if (i) the biological product is biosimilar to the reference product; and (ii) it can be expected to produce the same clinical result in *any given patient*. In addition, for a biological product that is administered more than once to an individual, the risk in terms of safety or diminished efficacy of alternating or switching between use of the biological product and the reference product is not greater than the risk of using the reference product without such alternation or switch.

Thus, there is a clear distinction between biosimilarity and interchangeability. In other words, biosimilarity does not imply interchangeability which is much more stringent. Intuitively, if a test product is judged to be interchangeable with the reference product then it may be substituted, even alternated, without a possible intervention, or even notification, of the healthcare provider. However, the interchangeability is expected to produce the *same* clinical result in *any given patient*, which can be interpreted as that the same clinical result can be expected in *every single patient*. In reality, conceivably, lawsuits may be filed if adverse effects are recorded in a patient after switching from one product to another.

It should be noted that when FDA declares the biosimilarity of two drug products, it may not be assumed that they are interchangeable. Therefore, labels ought to state whether for a follow-on biologic which is biosimilar to a reference product, interchangeability has or has not been established. However, payers and physicians may, in some cases, switch products even if interchangeability has not been established.

14.4.2 Switching and Alternation

Unlike drug interchangeability in terms of prescribability and switchability for generic drug products (Chow and Liu 2008), the US FDA has slightly perception of drug interchangeability for biosimilars. From the FDA's perspectives, interchangeability includes the concept of switching and alternating between an innovative biologic product (R) and its follow-on biologics (T). The concept of switching is referred to as not only the switch from "R to T" or "T to R" (narrow sense of switchability), but also "T to T" and "R to R" (broader sense of switchability). Note "T to T" could indicate a switch from an approved biosimilar product to another approved biosimilar product, while "R to R" could be a switch from an innovative biological product to itself (e.g., from a different batch or made at a different location). As a result, in order to assess switching, biosimilarity for "R to T," "T to R," "T to T," and "R to R" need to be assessed based on some biosimilarity criteria under a valid study design. The BPCI Act indicates that the risk in terms of safety or diminished efficacy of switching between use of the biological product and the reference product should not be greater than the risk of using the reference product without such a switch. This suggests that the risk of switching between T_i, $i = 1, \ldots, K$, where K is the number of approved biosimilars and R should not be greater than the risk of switching between R and R.

On the other hand, the concept of alternating is referred to as either the switch from T to R and then switch back to T (i.e., "T to R to T") or the switch from R to T and

then switch back to R (i.e., "R to T to R"). Thus, the difference between "the switch from T to R" then "the switch from R to T" and "the switch from R to T" then "the switch from T to R" needs to be assessed for addressing the concept of alternating. The BPCI Act also indicates that the risk in terms of safety or diminished efficacy of alternating between use of the biological product and the reference product should not be greater than the risk of using the reference product without such alternating. In practice, alternating "T to R to T" or "R to T to R" could have multiple T's (e.g., different approved biosimilars) and multiple R's (e.g., from different batches and/or made at different manufacturing locations/sites).

Thus, in practice, it is very difficult, if not impossible, to assess drug interchangeability of approved biosimilar products especially when there are multiple T's and R's in the marketplace. As stated in the BPCI Act, the relative risk between switching/alternating and without switching/alternating must be evaluated. However, little or no discussion about the criteria for assessment of the relative risk was mentioned in the BPCI Act. In the recent FDA draft guidances on the demonstration of biosimilarity of follow-on biologics, little or no discussion regarding the criteria, study design, and statistical methods for assessment of drug interchangeability in terms of switching and alternating was mentioned either. Thus, detailed regulatory guidances regarding the assessment of drug interchangeability in terms of switching and/or alternating are necessarily development.

14.4.3 Scaled Criteria for Drug Interchangeability (SCDI)

While the criterion for the determination of average biosimilarity is based on the BE requirement of (80.00, 125.00%) using log-transformed data, the criterion for the assessment of interchangeability is not clear. As indicated by Chen et al. (2000), variability due to subject-by-drug formulation interaction will have an impact on drug interchangeability. Let σ_D^2 be the variance component due to the subject-by-formulation interaction. We would like to propose a criterion to adjust for both intra-subject variability and the variability due to subject-by-formulation variability. The current BE criterion adjusted for intra-subject variability leads to so-called scaled average bioequivalence (SABE) criterion, which is considered suitable for highly variable drug products. In addition to adjusting intra-subject variability, following the idea of individual bioequivalence, we may further adjust the BE criterion with respect to the variability due to subject-by-formulation variability in order to have a more accurate and reliable assessment of interchangeability. As indicated in the 2001 FDA guidance on bioequivalence, criterion for assessing individual bioequivalence (IBE) is given by:

$$\theta = \frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + \left(\sigma_{WR}^2 - \sigma_{WT}^2\right)}{\max\left(\sigma_0^2, \sigma_{WR}^2\right)},$$

where σ_{WR}^2 and σ_{WT}^2 are intra-subject variances for the reference product and the test product, respectively, and σ_0^2 is a regulatory constant.

The proposed criterion will be based on the first two components of the criterion for individual bioequivalence, which consists of (i) criterion for average biosimilarity adjusted for intra-subject variability of the reference product (i.e., SABE), and (ii) correction for variability due to subject-by-product variability (i.e., σ_D^2). The proposed criterion for assessing interchangeability (i.e., switching and alternating) is briefly derived below.

Step 1: *Unscaled ABE criterion*

Let BEL be the BE limit which generally equals 1.25. Thus, biosimilarity requires that

$$\frac{1}{BEL} \leq GMR \leq BEL$$

This implies

$$-\log(BEL) \leq \log(GMR) \leq log(BEL),$$

or

$$-\log(BEL) \leq \mu_T - \mu_R \leq log(BEL),$$

where μ_T *and* μ_R are logarithmic means.

Step 2: *Scaled ABE* (*SABE*) *criterion*

Difference in logarithmic means is adjusted for intra-subject variability as follows:

$$-\log(BELS) \leq \frac{\mu_T - \mu_R}{\sigma_W} \leq \log(BELS),$$

or

$$-\log(BELS)\sigma_W \leq \mu_T - \mu_R \leq log(BELS)\sigma_W,$$

where σ_W^2 is a within-subject variation and *BELS* is the BE limit for SABE. In practice, σ_{WR}^2, the within-subject variation of the reference product is often considered.

Step 3: *Proposed scaled criterion for drug interchangeability* (*SCDI*)

Consider the first two components of the individual bioequivalence criterion, we have the following relationship:

$$\frac{(\mu_T - \mu_R)^2 + \sigma_D^2}{\sigma_W^2} = \frac{2\delta\sigma_D + (\delta - \sigma_D)^2}{\sigma_W^2},$$

where $\delta = \mu_T - \mu_R$. When δ and σ_D are close, we observe that

$$\frac{\delta^2 + \sigma_D^2}{\sigma_W^2} \approx \frac{2\delta\sigma_D}{\sigma_W^2}.$$

The assumption is reasonable when both δ and σ_D are small.

Thus, the proposed scaled criterion for drug interchangeability (SCDI) is:

$$-\log(BELS) \leq \left(\frac{\mu_T - \mu_R}{\sigma_W}\right)\left(\frac{2\sigma_D}{\sigma_W}\right) \leq \log(BELS).$$

Now, let $f = \sigma_W/(2\sigma_D)$, a correction factor for drug interchangeability. Then, the proposed SCDI criterion is given by:

$$-\log(BELS) f\sigma_W \leq \mu_T - \mu_R \leq log(BELS) f\sigma_W$$

Note that statistical properties and finite sample performance need further research.

Following the concept of criterion for individual bioequivalence and the idea of SABE, the proposed SCDI criterion is developed in order to adjust the usual one-size-fits-all approach for both intra-subject variability of the reference product and the variability due to subject-by-product. As compared to SABE, SCDI may result in wider or narrower limits depending upon the correction factor f which is a measure of the relative magnitude between σ_{WR} and σ_D.

The proposed SCDI criterion for drug interchangeability depends upon the selection of regulatory constants for σ_{WR} and σ_D. In practice, the observed variabilities may deviate far from the regulatory constants. Thus, it is suggested that the following hypotheses be tested before the use of SCDI criterion:

$$H_{01}: \sigma_{WR} > \sigma_{W0}\, vs.\, H_{a1}: \sigma_{WR} \leq \sigma_{W0},$$

and

$$H_{02}: \sigma_D > \sigma_{D0}\, vs.\, H_{a2}: \sigma_D \leq \sigma_{D0}$$

If we fail to reject the null hypotheses H_{01} or H_{02}, then we will stick with the suggested individual regulatory constants; otherwise, estimates of σ_{WR} and/or σ_D should be used in the SCDI criterion.

However, it should be noted that statistical properties and/or the finite sample performance of SCDI with estimates of σ_{WR} and/or σ_D are not well established. Further research is needed.

14.4.4 Study Designs

For assessment of bioequivalence for chemical drug products, a standard two-sequence, two-period (2×2) crossover design is often considered, except for drug products with relatively long half-lives. Since most biosimilar products have relatively long half-lives, it is suggested that a parallel group design should be considered. However, parallel group design does not provide independent estimates of variance components such as inter- and intra-subject variabilities and variability due to subject-by-product interaction. Thus, it is a major challenge for assessing biosimilarity (especially for assessing drug interchangeability) under parallel group designs since each subject will receive the same product once.

As indicated in the BPCI Act, for a biological product that is administered more than once to an individual, the risk in terms of safety or diminished efficacy of alternating or switching between use of the biological product and the reference product is not greater than the risk of using the reference product without such alternation or switch. Thus, for assessing drug interchangeability, an appropriate study design should be chosen in order to address (1) the risk in terms of safety or diminished efficacy of alternating or switching between use of the biological product and the reference product, (2) the risk of using the reference product without such alternation or switch, and (3) the relative risk between switching/alternating and without switching/alternating. In this section, several useful designs for addressing switching and alternation of biosimilar products are discussed.

Designs for Switching—Consider the broader sense of switchability. In this case, the concept of switching includes (1) switch from "R to T," (2) switch from "T to R," (3) switch from "T to T," and (4) switch to "R to R." Thus, in order to assess interchangeability of switching, a valid study design should be able to assess biosimilarity between "R and T," "T and R," "T and T," and "R and R" based on some biosimilarity criteria. For this purpose, the following study designs are useful.

Balaam design is a 4×2 crossover design, denoted by (TT, RR, TR, RT). Under a 4×2 Balaam's design, qualified subjects will be randomly assigned to receive one of the four sequences of treatments: TT, RR, TR, and RT. For example, subjects in sequence 1 of TT will receive the test (biosimilar) product first and then cross-overed to receive the reference (innovative biological) product after a sufficient length of washout (see Table 11.1). In practice, a Balaam design is considered the combination of a parallel design (the first two sequences) and a crossover design (sequences #3 and #4). The purpose of the part of parallel design is to obtain independent estimates of intra-subject variabilities for the test product and the reference product. In the interest of assigning more subjects to the crossover phase, an unequal treatment assignment is usually employed. For example, we may consider a 1:2 allocation to the parallel phase and the crossover phase. In this case, for a sample size of $N = 24$, 8 subjects will be assigned to the parallel phase and 16 subjects will be assigned to the crossover phase. As a result, 4 subjects will be assigned to sequences #1 and #2, while 8 subjects will be assigned to sequence #3 and #4 assuming that there is a 1:1 ratio treatment allocation within each phase.

Under Balaam's design, the first sequence provides not only independent estimate of the intra-subject variability of the test product, but also the assessment for "switch from T to T," while the second sequence provides independent estimate of the intra-subject variability of the reference product and compares difference between "R and R." The other two sequences assess similarity for "switch from T to R" and "switch from R to T," respectively. Under the 4×2 Balaam design, the following comparisons are usually assessed:

(1) Comparisons by sequence;
(2) Comparisons by period;
(3) T versus R based on sequence #3 and #4—this is equivalent to a typical 2×2 crossover design;
(4) T versus R given T based on sequence #1 and #3;
(5) R versus T given R based on sequence #2 and #4;
(6) The comparison between (1) and (3) for assessment of treatment-by-period interaction.

It should be noted that the interpretations of the above comparisons are different. More information regarding statistical methods for data analysis of Balaam design can be found in Chow and Liu (2008).

Designs for Alternating—For addressing the concept of alternating, an appropriate study design should allow the assessment of differences between "R to T" and "T to R" for alternating of "R to T to R" to determine whether the drug effect has returned to the baseline after the second switch. For this purpose, the following study designs are useful.

Two-sequence dual design is a 2×3 higher-order crossover design consisting of two dual sequences, namely TRT and RTR. Under the two-sequence dual design, qualified subjects will be randomly assigned to receive either the sequence of TRT or the sequence of RTR. Of course, there is a sufficient length of washout between dosing periods. Under the two-sequence dual design, we will be able to evaluate the relative risk of alternating between use of the biological product and the reference product and the risk of using the reference product without such alternating.

Note that expected values of the sequence-by-period means, analysis of variance table, and statistical methods (e.g., the assessment of average biosimilarity, inference on carryover effect, and the assessment of intra-subject variabilities) for analysis of data collected from a two-sequence dual design are given in Chow and Liu (2008). In case there are missing data (i.e., incomplete data), statistical methods proposed by Chow (2013) are useful.

For a broader sense of alternation involving more than two biologics, e.g., two biosimilars T_1 and T_2 and one innovative product R, there are six possible sequences: ($R\ T_2\ T_1$), ($T_1\ R\ T_2$), ($T_2\ T_1\ R$), ($T_1\ T_2\ R$), ($T_2\ R\ T_1$), and ($R\ T_1\ T_2$). In this case, a 6×3 William's design for comparing three products is useful (see, also, Chow and Liu 2008). A William design is a variance-balanced design, which consists of six sequences and three periods. Under the 6×3 William's design, qualified subjects

are randomly assigned to receive one of the six sequences. Within each sequence, a sufficient length of wash is applied between dosing periods.

Detailed information regarding (1) construction of a William design, (2) analysis of variance table, and (3) statistical methods for analysis of data collected from a 6×3 William design adjusted for carryover effects, in absence of unequal carryover effects, and adjusted for drug effect can be found in Chow and Liu (2008).

Designs for Switching/Alternating—In the previous two sub-sections, useful study designs for addressing switching and alternating of drug interchangeability are discussed, respectively. In practice, however, it is of interest to have a study design which can address both switching and alternating. In this case, an intuitive study design is to combine a switching design with an alternating design. Along this line, in this section, several useful designs for addressing both switching and alternating of drug interchangeability are introduced.

As indicated earlier, Balaam's design is useful for addressing switching, while a two-sequence dual design is appropriate for addressing alternating. In the interest of addressing both switching and alternating in a single trial, we may combine the two study designs as follows: (TT, RR, TRT, RTR), which consists of a parallel design (the first two sequences) and a two-sequence dual design (the last two sequences). Data collected from the first two dosing periods (which are identical to the Balaam design) can be used to address switching, while data collected from sequences #3 and #4 can be used to assess the relative risks of alternating.

As it can be seen that the modified Balaam's design in not a balanced design in terms of the number of dosing periods. In the interest of balance in dosing periods, it is suggested the modified Balaam's design be further modified as (TTT, RRR, TRT, RTR). We will refer to this design as a complete design. The difference between the complete design and the modified Balaam design is that the treatments are repeated at the third dosing period for sequences #1 and #2. Data collected from sequence #1 will provide a more accurate and reliable assessment of intra-subject variability, while data collected from sequence #2 is useful in establishing baseline for the reference product. Note that statistical methods for analysis of data collected from the complete design are similar to those under the modified Balaam's design.

14.4.5 Unified Approach for Assessing Interchangeability

In practice, switching and alternating can only be assessed after the biosimilar products under study have been shown to be highly similar to the innovative biological drug product. Based on similar idea for development biosimilarity index (Chow et al. 2011), a general approach for development of switching index and/or alternating index for addressing switching and/or alternating can be obtained.

Switching Index (SI)—Similar idea can be applied to develop switching index under an appropriate study design such as a 4×2 Balaam's crossover design described

earlier. Thus, biosimilarity for "R to T," "T to R," "T to T," and "R to R" need to be assessed for addressing the issue of switching.

Define $\hat{p}_{Ti}$ the totality biosimilarity index for the ith switch, where $i=1$ (switch from R to R), 2 (switch from T to T), 3 (switch from R to T), and 4 (switch from T to R). As a result, the switching index (SI) can be obtained as follows:

Step 1: Obtain $\hat{p}_{Ti}$, $i=1, \ldots, 4$;
Step 2: Define switching index as $SI = \min_i\{\hat{p}_{Ti}\}$, $i=1, \ldots, 4$, which is the largest order of the biosimilarity indices;
Step 3: Claim switchability if the 95% confidence lower bound of $\min_i\{p_{Ti}\}$. $i=1, \ldots, 4$, is greater than a pre-specified value p_{s0}.

Alternating Index (**AI**)—Similar idea can be applied to develop alternating index under an appropriate study design. Under the modified Balaam's crossover design of (TT, RR, TRT, RTR), biosimilarity for "R to T to R" and "T to R to T" need to be assessed for evaluation of alternating. For example, the assessment of differences between "R to T" and "T to R" for alternating of "R to T to R" need to be evaluated in order to determine whether the drug effect has returned to the baseline after the second switch.

Define p_{Ti} the totality biosimilarity index for the ith switch, where $i=1$ (switch from R to R), 2 (switch from T to T), 3 (switch from R to T), and 4 (switch from T to R). As a result, the alternating index (AI) can be obtained as follows:

Step 1: Obtain $\hat{p}_{Ti}$, $i=1, \ldots, 4$;
Step 2: Define the range of these indexes, $AI = \max_i\{\hat{p}_{Ti}\text{QUOTE}\} - \min_i\{\hat{p}_{Ti}\text{QUOTE}\}$, $i = 1, \ldots, 4$, as the alternating index;
Step 3: Claim alternation if the 95% confidence lower bound of $\max_i\{p_{Ti}\} - \min_i\{p_{Ti}\}$, $i=1, \ldots, 4$, is greater than a pre-specified value p_{A_0}.

14.5 Concluding Remarks

14.5.1 Biosimilarity Assessment

Biological products or medicines are therapeutic agents made of a living system or organism. As a number of biologic products will be due to expire in the next few years, the potential opportunity in developing the follow-on products of these originator products may result in the reduction of these products and provide more choices to medical doctors and patients for getting the similar treatment care with lower cost. However, the price reductions versus the originator biologic products remain to be determined, as the advantage of a slightly cheaper price may be outweighed by the hypothetical increased risk of side-effects from biosimilar molecules that are not exact copies of their originators. Unlike traditional small molecule drug products, the characteristic and development of biologic products are more complicated and

sensitive to many factors. Any small change in manufacturing process may result in the change of therapeutic effect of the biologic products. The traditional bioequivalence criterion for average bioequivalence of small molecule drug products may not be suitable for evaluation of biosimilarity of biologic products. Therefore, in this article, we evaluate the biosimilar index proposed by Chow et al. (2011) for assessment of the (average) biosimilarity between innovator and reference products. Both results based on estimation and Bayesian approaches demonstrate that the proposed method based on biosimilar index can reflect the characteristics and impact of variability on the therapeutic effect of biologic products. However, the estimated reproducibility probability based on the Bayesian approach depends on the choice of the prior distributions. If a different prior such as an informative prior is used, a sensitivity analysis may be performed to evaluate the effects of different prior distributions.

The other advantage of the proposed method can be applied to different functional areas (domains) of biological products such as pharmacokinetics (PK), biological activities, biomarkers (e.g., pharmacodynamics), immunogenicity, manufacturing process, efficacy, etc., since it is developed based on the probability of reproducibility. The further research will be employed for the development of the statistical testing approach for the evaluation of biosimilarity across domains.

Current methods for the assessment of bioequivalence for drug products with identical active ingredients are not applicable to follow-on biologics due to fundamental differences. The assessment of biosimilarity between follow-on biologics and innovator in terms of surrogate endpoints (e.g., pharmacokinetic parameters and/or pharmacodynamics responses) or biomarkers (e.g., genomic markers) requires the establishment of the fundamental biosimilarity assumption in order to bridge the surrogate endpoints and/or biomarker data to clinical safety and efficacy.

Unlike conventional drug products, follow-on biologics are very sensitive to small changes in variation during the manufacturing process, which have been shown to have an impact on the clinical outcome. Thus, it is a concern whether current criteria and regulatory requirements for the assessment of bioequivalence for drugs with small molecules can be applied also to the assessment of biosimilarity of follow-on biologics. It is suggested that current, existing criteria for the evaluation of bioequivalence, similarity, and biosimilarity be scientifically/statistically evaluated in order to choose the most appropriate approach for assessing biosimilarity of follow-on biologics. It is recommended that the selected biosimilarity criteria should be able to address (1) sensitivity due to small variations in both location (bias) and scale (variability) parameters, and (2) the degree of similarity, which can reflect the assurance for drug interchangeability.

Under the established fundamental biosimilarity assumption and the selected biosimilarity criteria, it is also recommended that appropriate statistical methods (e.g., comparing distributions and the development of biosimilarity index) be developed under valid study designs (e.g., Design A and Design B described earlier) for achieving the study objectives (e.g., the establishment of biosimilarity at specific domains or drug interchangeability) with a desired statistical inference (e.g., power or confidence interval). To ensure the success of studies conducted for the assessment of biosimilarity of follow-on biologics, regulatory guidelines/guidances need to be

developed. Product-specific guidelines/guidances published by EU EMA have been criticized for not having standards. Although product-specific guidelines/guidances do not help to establish standards for the assessment of biosimilarity of follow-on biologics, they do provide the opportunity for accumulating valuable experience/information for establishing standards in the future. Thus, We recommends that several numerical studies are recommended including simulations, meta-analysis, and/or sensitivity analysis, in order to (1) provide a better understanding of these product-specific guidelines/guidances, and (2) check the validity of the established Fundamental Biosimilarity Assumption, which is the legal basis for assessing biosimilarity of follow-on biologics.

14.5.2 Analytical Similarity Assessment

For identifying CQAs at various stages of the manufacturing process, most sponsors assign CQAs based on the mechanism of action (MOA) or pharmacokinetics (PK) which are believed to be relevant to clinical outcomes. It is a reasonable assumption that change in MOA or PK of a given quality attribute is predictive of clinical outcomes. However, the primary assumption that there is a well-established relationship between in vitro assays and in vivo testing (i.e., in vitro assays and in vivo testing correlation; IVIVC) needs to be validated. Under the validated IVIVC relationship, the criticality (or risk ranking) can then be assessed based on the degree of the relationship. In practice, however, most sponsors provide clinical rationales for the assignment of the CQAs without using a statistical approach for the establishment of IVIVC. The assignment of the CQAs without using a statistical approach is considered subjective and hence is somewhat misleading.

For a given quality attribute, FDA suggests a simple approach by testing one sample (randomly selected) from each of the lots. Basically, FDA's approach ignores lot-to-lot variability for the reference product. In practice, however, lot-to-lot variability inevitably exists even when the manufacturing process has been validated. In other words, we would expect that there are differences in mean and variability from lot-to-lot, i.e., $\mu_{Ri} \neq \mu_{Rj}$ and $\sigma^2_{Ri} \neq \sigma^2_{Rj}$ for $i \neq j, i, j = 1, 2, \ldots, K$. In this case, it is suggested that FDA's approach be modified (e.g., performing tests on multiple samples from each lot) in order to account for the within-lot and between-lot (lot-to-lot) variabilities for fair and reliable comparisons.

For the quality range approach for CQAs in Tier 2, FDA recommends to use $x = 3$ by default for 90% of values of test lots contained in the range. It allows approximately one standard deviation of reference for shifting, which may be adjusted based on biologist reviewers' recommendations. However, some sponsors propose using the concept of tolerance interval in order to ensure that there is a high percentage of test values for the lots from the test product fall within the quality range. It, however, should be noted that the percentage decreases when the difference in mean between the reference product and the proposed biosimilar product increases. This is also true when $\sigma_T \ll \sigma_R$. Even the tolerance interval is used as the quality range. This

problem is commonly encountered mainly because the quality range approach does not take into consideration (i) the difference in means between the reference product and the proposed biosimilar product, and (ii) the heterogeneity among lots within and between products. In practice, it is very likely that a biosimilar product with small variability but a mean response which is away from the reference mean (e.g., within the acceptance range of $\sigma_R/8$ per FDA) will fall outside the quality range. In this case, a further evaluation of the data points that fall outside the quality range is necessary to rule out the possibility by chance alone.

FDA's current thinking for analytical similarity assessment using a 3-tier analysis is encouraging. It provides a direction for statistical methodology development for a valid and reliable assessment toward providing the totality-of-the-evidence for demonstrating biosimilarity. The 3-tier approach is currently under tremendous discussion within the pharmaceutical industry and academia. In addition to the challenging issues discussed above, there are some issues that remain unsolved and require further research. These issues include, but are not limited to, (i) the degree of similarity (i.e., how similar is considered highly similar?), (ii) multiplicity (i.e., is there a need to adjust α for controlling the overall type I error at a pre-specified level of significance), (iii) acceptance criteria (e.g., about what percentage of CQAs in Tier 1 need to pass an equivalence test in order to pass the analytical similarity test for Tier 1?), (iv) multiple references (i.e., what if there are two reference products such as US-licensed and EU-approved reference product), and (v) credibility toward the totality-of-the-evidence.

14.5.3 Assessing Drug Interchangeability

With small molecule drug products, bioequivalence generally reflects therapeutic equivalence. Drug prescribability, switching, and alternating are generally considered reasonable. With biologic products, however, variations are often higher (other than pharmacokinetic factors may be sensitive to small changes in conditions). Thus, often only parallel group design rather than crossover kinetic studies can be performed. It should be noted that very often, with follow-on biologics, biosimilarity does not reflect therapeutic comparability. Therefore, switching and alternating should be pursued with extremely caution.

The concept of drug interchangeability in terms of prescribability and switchability for small molecule drug products are similar but different from those for large molecule biological products as defined in the BPCI Act. Thus, the usual methods for addressing drug interchangeability through the assessment of population/individual bioequivalence cannot be directly applied for assessment of drug interchangeability for biosimilar products. For biosimilar products, the assessment of drug interchangeability in terms of the concepts of switching and alternating involves (1) the assessment of biosimilarity and (2) the evaluation of the relative risk of switching and alternating. It should be noted that there is a clear distinction between the assessment

of biosimilarity and the evaluation of drug interchangeability. In other words, the demonstration of biosimilarity does not imply drug interchangeability.

Based on totality biosimilarity index for assessment of biosimilarity, switching index and alternating index for addressing drug interchangeability of biosimilar products can be obtained under appropriate switching design and alternating design, respectively. The proposed switching and alternating indices have the advantages that (1) they can be applied regardless of the criteria for biosimilarity and study design used, (2) the assessment is made based on relative difference with the reference product (i.e., relative difference between (T vs. R) and (R vs. R), (3) it can address the commonly asked question that "how similar is considered highly similar?", "the degree of similarity," and "interchangeability in terms of switching and alternating" in terms of the degree of reproducibility, and most importantly (4) the proposed method is in compliance with current regulatory thinking (i.e., totality-of-the-evidence, relative risk of switching and alternating for interchangeability). It, however, should be noted that the proposed totality biosimilarity index and/or switching and alternating indices depend upon the selection of weights in each domain or functional area for achieving the totality-of-the-evidence for assessment of biosimilarity and/or interchangeability. The performances of the proposed totality biosimilarity index, switching index, and alternating index are currently being studied via clinical trial simulations by Starr et al. (2013).

In practice, it is a concern regarding how many subjects are needed for providing totality-of-the-evidence across different functional areas such as PK/PD, clinical efficacy/safety, and manufacturing process and addressing relative risks of switching and alternating for interchangeability. Based on the proposed totality biosimilarity index (for assessment of biosimilarity) and switching and alternating indices (for assessment of interchangeability), sample size required for achieving certain statistical inference (assurance) can be obtained following the procedure as described in Chow (2013).

As indicated earlier, the broader sense of the concepts of switching and alternating could involve in a number of biosimilar products, e.g., T_i, i = 1, …, K, which has been shown to be biosimilar to the same innovative (reference) drug product. Under the broader sense of interchangeability, it is almost impossible for a sponsor to claim or demonstrate interchangeability according to the definition as given in the BPCI Act. Alternatively, it is suggested that a meta-analysis that combines all of data given in the submissions be conducted (by the regulatory agency) to evaluate the relative risks of switching and alternating of drug interchangeability. In practice, meta-analysis can be conducted for safety monitoring of approved biosimilar products. This is extremely important especially when there are a number of biosimilar products in the marketplace.

References

Chen, M. L., Patnaik, R., Hauck, W. W., Schuirmann, D. F., Hyslop, T., & Williams, R. (2000). An individual bioequivalence criterion—Regulatory considerations. *Statistics in Medicine, 19,* 2821–2842.

Chow, S. C. (2013). *Biosimilars: Design and analysis of follow-on biologics*. Chapman and Hall/CRC Press, Taylor & Francis, New York.

Chow, S. C. (2014). On assessment of analytical similarity in biosimilar studies. *Drug Designing, 3,* 119. https://doi.org/10.4172/2169-0138.

Chow, S. C. (2015). Challenging issues in assessing analytical similarity in biosimilar studies. *Biosimilars, 5,* 33–39.

Chow, S. C., & Liu, J. P. (2008). *Design and analysis of bioavailability and bioequivalence studies* (3rd ed). New York, New York: Chapman Hall/CRC Press, Taylor & Francis

Chow, S. C., Hsieh, T. C., Chi, E., & Yang, J. (2010). A comparison of moment-based and probability-based criteria for assessment of follow-on biologics. *Journal of Biopharmaceutical Statistics, 20,* 31–45.

Chow, S. C., Endrenyi, L., Lachenbruch, P. A., Yang, L. Y., & Chi, E. (2011). Scientific factors for assessing biosimilarity and drug interchangeability of follow-on biologics. *Biosimilars, 1,* 13–26.

Chow, S. C., Yang, L. Y., Starr, A., & Chiu, S. T. (2013). Statistical methods for assessing interchangeability of biosimilars. *Statistics in Medicine, 32,* 442–448.

Christl, L. (2015) Overview of the regulatory pathway and FDA's guidance for the development and approval of biosimilar products in the US. Presented at the *Oncologic Drugs Advisory Committee meeting*, January 7, 2015, Silver Spring, Maryland.

Davit, B. M., Conner, D. P., Fabian-Fritsch, B., Haidar, S. H., Jiang, X., Patel D. T., et al. (2008). Highly variable drugs: Observations from bioequivalence data submitted to the FDA for new generic drug applications. *AAPS Journal, 10*, 148–156.

FDA. (2003). Guidance on bioavailability and bioequivalence studies for orally administrated drug products—General considerations. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland.

FDA. (2012a). Scientific considerations in demonstrating biosimilarity to a reference product. The United States Food and Drug Administration, Silver Spring, Maryland, USA.

FDA. (2012b). Quality considerations in demonstrating biosimilarity to a reference protein product. The United States Food and Drug Administration, Silver Spring, Maryland, USA.

FDA. (2012c). Biosimilars: Questions and Answers Regarding Implementation of the Biologics Price Competition and Innovation Act of 2009. The United States Food and Drug Administration, Silver Spring, Maryland, USA.

FDA. (2015). Guidance for industry: Scientific considerations in demonstrating biosimilarity to a reference product. Food and Drug Administration, Silver Spring, Maryland.

Haidar, S. H., Davit, B., Chen, M. L., Conner, D., Lee, L., Li, Q. H., et al. (2008a). Bioequivalence approaches for highly variable drugs and drug products. *Pharmaceutical Research, 25,* 237–241.

Haidar, S. H., Makhlouf, F., Schuirmann, D. J., Hyslop, T., Davit, B., Conner, D., et al. (2008b). Evaluation of a scaling approach for the bioequivalence of highly variable drugs. *The AAPS Journal, 2008*(10), 450–454.

Hsieh, T. C., Chow, S. C., Liu, J. P., Hsiao, C. F., & Chi, E. (2010). Statistical test for evaluation of biosimilarity of follow-on biologics. *Journal of Biopharmaceutical Statistics, 20*(20), 75–89.

Schuirmann, D. J. (1981). On hypothesis testing to determine if the mean of a normal distribution is continued in a known interval. *Biometrics, 37,* 617.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15,* 657–680.

Tothfalusi, L., Endrenyi, L., & Garcia, A. A. (2009). Evaluation of bioequivalence for highly-variable drugs with scaled average bioequivalence. *Clinical Pharmacokinetics, 48,* 725–743.

Tsong, Y. (2015) Development of statistical approaches for analytical biosimilarity evaluation. in *DIA/FDA Statistics Forum 2015*, April 20, 2015, Bethesda, Maryland.

Yu., L. X. (2004). Bioinequivalence: concept and definition. Presented at *Advisory Committee for Pharmaceutical Science of the Food and Drug Administration*, 13–14 April, 2004, Rockville, Maryland.

Chapter 15
Causal Estimands: A Common Language for Missing Data

Steven A. Gilbert and Ye Tan

Incorrect choice of estimand and unclear definitions for estimands lead to problems in relation to trial design, conduct and analysis and introduce potential for inconsistencies in inference and decision making. (ICH Concept Paper 2014)

15.1 Introduction

What are estimands, and what is there connection with missing data? The topic of estimands is all about interpreting the results of our statistical analyses in the context of the original scientific questions that we intend to answer in a clinical trial. Missing data, a ubiquitous problem in clinical trials, require additional assumptions and choices on how to collect and analyze data that can have an unexpected impact on the interpretation of these analyses. When taking an 'estimand approach' to planning a clinical trial, we plan ahead for missing data, protocol deviations, and other unwanted events so that the interpretation and context of our statistical results are maintained. We hope to impress upon statisticians that at its core, the estimand approach is not new, but rather a clarification of good statistical practice, taking assumptions implicit in our analyses and making them explicit. Clinical trials are a scientific endeavor, and statisticians must be a fully functioning member of the scientific process; our analyses need to incorporate the best scientific knowledge available which we can only obtain by communicating with our colleagues.

S. A. Gilbert (✉) · Y. Tan
Early Clinical Development, Pfizer Inc., Cambridge, MA, USA
e-mail: Steven.A.Gilbert@Pfizer.com

Y. Tan
e-mail: Ye.Tan@Pfizer.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_15

A very basic model of a statistical analysis goes as follows. A sample is drawn from a *population*, and a statistic is calculated, say the mean for example. Knowing the mean of that particular sample drawn is usually not of great interest, but knowing the mean of the population the sample was drawn from is of interest. Statistical theory can tell us about how well the sample mean approximates the population mean. The key to estimands is understanding what population we are talking about.

Regression models are a common statistical tool in all types of research including clinical trials. A regression analysis will provide estimates of β coefficients that may be interpreted as a treatment effect or effect of a baseline measurement such as age or disease severity. In a mathematical statistics class, you assume there is a true β and the statistical method produces an estimate $\hat{\beta}$. You can then go on to prove consistency, optimality, and other mathematical measures of how well $\hat{\beta}$ approximates β. The underlying assumption is that the true β is well defined and easy to identify. Or equivalently the population of subjects that are described by these regression coefficients is easily identified. In a clinical trial, especially when subjects withdraw early, violate the protocol, and do not provide all expected data, considerable thought may be needed to identify the 'true' underlying parameter, or equivalently, what larger population of subjects do the βs describe?

To illustrate the difficulty in identifying the underlying estimand, let us look at a classical example from the physical sciences literature and contrast it to a hypothetical clinical trial. The physical science example is from an experiment by Newcomb and Michaelson in September of 1882 when they measured the speed of light by repeatedly timing how long it took for light to travel from their laboratory on the Potomac River to the base of the Washington Monument and back, a distance of approximately 7400 m (MacKay and Oldford 2000). The time measurements can be transformed to speed, and the average speed is interpreted as the speed of light or at least the speed of light in the atmosphere. In contrast, the situation is much more complex in clinical trials. Consider a clinical trial with a new GLP-1 receptor antagonist to control blood sugar in diabetics as measured by change from baseline HbA_{1c} at 26 weeks. Over the course of the 26-week study, many subjects complete the study as planned; however, a significant number of subjects are: lost to follow-up, require rescue medications not specified in the protocol, have an adverse reaction to the treatment, or refuse to take their medication as directed. The available data are then analyzed as specified in the statistical analysis plan (SAP), for example, analysis of covariance, with missing values imputed using last observation carried forward. Finally, the results are tabulated, and the treated group reduces their HbA_{1c} by 1.12% versus 0.10% for the control arm, with a difference of 1.02 favoring the new treatment. What does this estimated difference mean? To what population and under what circumstances can these results be generalized? Is it all diabetics? Do they need to be completely compliant with dosing? What if they are on a background therapy or if they are over 80 years of age? What is the target of our inference?

The estimated treatment difference above is a long way from obtaining data measuring a fundamental physical constant that ostensibly is the same anywhere and at any time in the universe. Trial results can depend upon many factors including

background therapy that differs from one geographical area to another and evolves over time, potentially making the results both time and location specific.

Another complication is that any long-term trial will have subjects experiencing different events after their initial randomization to study drug. These post-randomization events range from the minor, missing an occasional dosing, to major, dropping from the study or switching to a rescue medication. Each of these post-randomization events changes the actual treatment received from study drug alone to study drug plus rescue medication or study drug at a lower dose than planned, etc. How these complications are accounted for in the statistical analysis (e.g., multiple imputation, last observation carried forward, complete case analysis) will effect how the results can be generalized. The estimand is merely a clear statement of how the results of the study can be generalized to a larger population, namely who is in that population and under what circumstances do the results hold.

To make this a little less abstract, consider the following examples of estimands taken from the 'Prevention and Treatment of Missing Data in Clinical Trials,' (NRC 2018).

1. (Difference in) Outcome improvement for all randomized participants.
2. (Difference in) Outcome improvement in tolerators.
3. (Difference in) Outcome improvement if all subjects tolerated or adhered.
4. (Difference in) Areas under the outcome curve during adherence to treatment.
5. (Difference in) Outcome improvement during adherence to treatment.

Later on, Mallinckrodt et al. 2012 suggested a sixth estimand.

6. Difference in outcome improvement in all randomized patients at the planned endpoint of the trial attributable to the initially randomized medication.

At a more fundamental level, there are two large classes of estimands for endpoints, effectiveness, and efficacy. Table 15.1 shows the two basic categories. On the left is effectiveness, which is more or less how well does the drug work under less than ideal conditions. The right side has efficacy, which is how well does the drug work when taken as directed that is ideal conditions. Regulators tend to prefer effectiveness, while sponsors prefer efficacy. The preference of clinicians will depend on the type of drug and the disease that is being treated. Drugs that are used for symptomatic relief, say analgesics for pain, are better characterized by efficacy. If a subject does not tolerate a drug or if it is ineffective, another drug can be tried. Vaccines or treatments

Table 15.1 Categories of efficacy estimands

Effectiveness	Efficacy
de facto	de jure
ITT	Per protocol
Treatment policy	Drug effect

for life-threatening diseases such as cancer, where there is less opportunity to try different treatments, may be better characterized by effectiveness.

The terms de facto and de jure are Latin terms proposed by Carpenter, Roger, and Kenward that roughly translate as 'by fact' and 'by the law,' and refer to effectiveness and efficacy. The benefit of this terminology is that it can be applied to safety analyses as well, though this terminology is not universally accepted. Similar concepts are expressed by the idea of using an intention to treat (ITT) or per protocol (PP) analysis. Lastly, these two general types of analyses can be viewed as assessing either the 'treatment policy' (i.e., what is the effect of prescribing the drug, whether you take it or not) and drug effect (i.e., what is the drug doing when properly dosed).

If subjects are completely compliant with the protocol, and in the absence of missing data, the estimand framework is easy to apply and may even seem superfluous. In the ideal case, all subjects comply with the protocol and all data are available. The results apply to the population of subjects described by the inclusion/exclusion criteria who follow the protocol. In the less than ideal case, post-randomization events complicate the interpretation of the estimand and subjects with missing observations present new challenges regarding how to include them at all in the analysis. We suggest that the estimand should reflect the interpretation of the trial results from the ideal situation. The next step is to choose statistical methods and possible alterations to the study design so that meaningful estimates can still be obtained under the less than ideal case.

Estimands are no more and no less than remembering to start and end with the science, a lesson well known in the statistics community. Start by understanding and defining the scientific questions that need to be answered. Choose the best statistical tools for the job, and interpret the statistical results back into science. In clinical trials, this translates to 'how does a treatment work in subjects?' and ends with an answer to that question. What follows are some useful methods to help make the underlying scientific question more clear to ourselves and our colleagues, and some methods to help keep from going astray on that simple course from the original scientific question to its eventual answer.

15.2 Theoretical Framework

A theoretical framework is presented to act as a bridge between the scientific goals of the study and the mathematical details of the statistical analyses. This will begin an iterative process of planning and designing that will most likely require simulation experiments to determine sample size, power, precision, and other important operating characteristics.

The theoretical framework to be described is not unique and is taken from the literature. The authors' only contribution is to highlight the areas that we think are useful for the majority of clinical trials that we are involved with, namely missing data factorizations and some basic causal inference. We intend to review just enough

theory to help in understanding estimands and how they can be implemented in practice omitting many details.

Beginning with missing data factorizations, we focus on pattern mixture models (PMMs), then describe the conditional predictive distribution (CPD), and finally review some causal inference topics at a very high level. The CPD describes our knowledge of the missing data. Since by definition we do not observe the missing data directly, our knowledge of that data is more uncertain than the observed data. Using a statistical distribution to describe the unseen data is a handy way to put both structures on the data (e.g., mean parameters in a normal distribution) and keep track of our uncertainty via the spread and variability of the distribution.

The CPD describes in a sense how missing data are 'filled in,' but this only takes us half way to our goal making a fair comparison of different treatments consistent with our desired estimand. This is where a little familiarity with the causal inference literature is important. The general idea is to make sure you are comparing like with like. The theory makes this more rigorous and provides a notation that greatly simplifies thinking about these issues.

15.2.1 Pattern Mixture and Selection Model Factorizations

Begin by defining the data $Y = \{Y_1, \ldots, Y_n\}$ where each Y_i is vector containing the complete data on the endpoint of interest for subject i. If the random variables are continuous, assume they will be modeled with a density from the exponential family (Pawitan 2001) of distributions. For example, consider univariate observations sampled independently from $Y \sim N(\theta, \sigma^2)$ with parameters θ and σ. Without loss of generality, consider σ known. The density function for Y is the product of the individual densities, $f(Y) = \prod f(y_i)$, where

$$f(y_i) = \left(2\pi\sigma^2\right)^{-1/2} exp^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}}.$$

The joint density is called the likelihood function $\mathcal{L}(\theta)$ when it is treated as a function of θ with the observations, $y_i's$, fixed at their observed values. The maximum likelihood estimate (MLE) is calculated by finding the parameter values that maximize[1] the likelihood function or equivalently the natural logarithm of the likelihood, the log-likelihood. Even in a simple example, θ will most likely be a vector. For example, a two-arm parallel trial comparing means at a single time point could use a model such as $E(y_i) = \beta_0 + \beta_D D_i$ where $D_i = 1$ if subject i is randomized to active drug and 0 otherwise. In this case, $\theta = \{\beta_0, \beta_D\}$[2] where the β_D is *interpreted* as the additive treatment effect of the drug on treated subjects. The topic of estimands is all

[1] We ignore pathological cases such as multiple maxima to keep the discussion simple.

[2] Nuisance parameters such as the variance are ignored for now to simplify the exposition.

about this last step, interpreting the results of our statistical analysis in the context of our original scientific question.

The longer a trial lasts, the greater the chance that we will not be able to measure a response at a desired time, landmark visit, on all subjects. They may refuse to return to the clinic, stop taking drug on their own, and have an adverse event where the investigator removes the drug for safety reasons, etc. The question now is, can we extend the simple probabilistic framework that was so useful for maximum likelihood estimation to this more complicated setting? The answer, fortunately, is yes.

The framework can be extended by modeling both the response data as above and the missing data mechanism as well. Define a response indicator $R = \{R_1, \ldots, R_n\}$ that takes values of 1 when data are observed and 0 when they are missing.[3] It will also be useful to add some additional notation to keep track of the observed and missing responses. Borrowing common notation from the missing data literature, 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 contain the treatment assignment and other covariates of interest.

The next step is to define the joint distribution of $\{Y_O, Y_M, R\}$. Simply defining a joint distribution for these three variables would be difficult or impossible in most situations. However, the joint distribution can be factored into conditional distributions that are more tractable.

The two most common factorizations are the selection model and pattern mixture factorizations, and mathematically they are described as:

$$P[Y_O, Y_M, R|X] = P[R|Y_O, Y_M, X] \times P[Y_O, Y_M|X] \tag{15.1}$$

a selection model and

$$P[Y_O, Y_M, R|X] = P[R|X] \times P[Y_O, Y_M, |R, X], \tag{15.2}$$

a pattern mixture model(Carpenter and Kenward 2012). A third factorization, using a shared latent parameter, can also be used but will not be discussed in this chapter.

The second term on the right-hand side (RHS) of Eq. 15.1 $P[Y_O, Y_M|X] = P[Y|X]$ is the distribution of the complete data. The first term on the RHS of Eq. 15.1, $P[R|Y_O, Y_M, X]$, describes the conditional probability of observing (or selecting) each element of the complete data vector; let us call that the selection distribution for future reference. This selection distribution is used to classify missing data mechanisms as: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), to be described below.

A good way to gain an intuitive understanding of $P[R|Y_O, Y_M, X]$ is think of modeling $P(R) = 1$ using logistic regression. If $P[R|Y_O, Y_M, X] = P[R]$, the logistic model would only have an intercept and the data are MCAR. In this case, the miss-

[3]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 PMMs.

ing data can be ignored and any method including simple summary statistics of the observed data will be unbiased.

If $P[R|Y_O, Y_M, X] = P[R|Y_O, X]$, the logistic regression model will only have covariates that are included in the observed data, and the missing data are then said to be MAR. That is a model for R can be built using the observed endpoints and covariates. If in addition to MAR the data and missing data distributions have functionally distinct parameters, (i.e., $P[Y_O, Y_M|X; \theta]$, $P[R|Y_O, Y_M, X, ; \eta]$ with θ and η functionally distinct), the missing data mechanism is said to be ignorable. In this case, only the response data, the $Y_i's$, need to be modeled to obtain an unbiased estimate, while a model for the $R_i's$ is not needed. The model for the response generally needs to use all of the data in the available repeated measures on each subject and to correctly model the dependencies over time. This is often done with longitudinal data where the claim is made that data are expected to be missing at random and therefore a mixed model repeated measures (MMRM) analysis will be unbiased. However, summary statistics and models that do not incorporate all of the relevant observed response and covariates will be biased.

Lastly, if $P[R|Y_O, Y_M, X]$ cannot be simplified further, the logistic regression model will require Y_M, and the data are said to be MNAR. This is the most difficult of the three missing data scenarios to handle.

We caution that categorizing the missing data as MCAR, MAR, or MNAR alone is not describing an estimand. Rather, it is an important part of defining and estimating an estimand in the presence of missing data.

15.2.2 Conditional Predictive Distribution

Continuing with the pattern mixture model factorization, the missing data can be integrated out to obtain the observed likelihood.

$$\begin{aligned} P[Y_O, R] &= \int P[Y_O, Y_M, R] dY_M \\ &= \int P[R] P[Y_O, Y_M|R] dY_M \\ &= \int P[R] P[Y_O|R] P[Y_M|Y_O, R] dY_M \end{aligned}$$

The term $P[Y_M|Y_0, R]$ is the distribution of the missing data conditional on the observed data and response indicators, sometimes referred to as the conditional predictive distribution (Carpenter and Kenward 2012). This is the key function for inference and the basis of multiple imputation (MI) approaches. Multiple imputation draws a random sample from the conditional predictive distribution, fits a model, and repeats the process multiple times keeping track of the results. These results are then combined into a single estimate using Rubin's rules for multiple imputation

(Carpenter and Kenward 2012). Clearly defining an estimand requires a clear definition of the conditional predictive distribution, which is really just a description of what the missing data are expected to look like. This is the main area where statisticians need the input of their clinical and scientific colleagues.

15.2.3 Causal Models

The last section covering selection models, pattern mixture models, and conditional predictive distributions addressed the issue of what the missing data look like and the uncertainty in its true value. We now bring in a few concepts from causal inference to help compare treatment groups. Gelman and Hill (2007) describe causal inference this way 'In the usual regression context, predictive inference relates to comparisons *between* units, whereas causal inference addresses comparisons of different treatment if applied to the *same* units,' where of course we can think of units as subjects. In most cases, it is impossible to apply the same treatment to the same subject, and in practice we infer what would happen to a subject based on comparing different subjects. For this chapter, we will refer to a comparison of treatments on different subjects that can be substituted for a comparison of different treatments within the same subject as a 'fair' comparison. We will simplify the exposition greatly and only concentrate on the main concepts germane to this chapter. The reader interested in a comprehensive review of causal models is encouraged to read 'Causal effect in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches,' published in 2000 in the Annual Review of Public Health, (Little and Rubin 2000).

Without loss of generality, consider two treatments, active drug and control. Define $y_i(D = d)$ as the outcome for subject i if they receive treatment $D = d$, for example $D = 1$ for active treatment and $D = 0$ for control. These are referred to as potential outcomes, since only one of the two will be observed depending upon the randomization assignment.

Table 15.2 is artificial but demonstrates some important points. In this framework, everyone has a response to treatment and control, but we only get to observe one or the other. This is referred to as the 'fundamental problem' of causal inference. Note that in this context, even a perfectly run trial where everyone stays on protocol and contributes data has 50% of the data missing!.

For now, let the table represent an entire population of six subjects. In this population, the treatment mean is 11 and the control mean response is 2. The treatment difference δ is the same for any value of the baseline covariate. Therefore, comparing the two subjects with a baseline value of 40 would yield the same $\hat{\delta} = 12 - 3$ of 9 as comparing the two subjects with a baseline value of 20, $\hat{\delta} = 10 - 1 = 9$. However, the average responses would differ; for example, the treated subjects have an average response of 12 when the baseline is 40 and only 10 when the baseline is 20. The estimation of δ can be biased if the probability of choosing subjects based on baseline covariate values differs in the treated and control arms. Comparing the

Table 15.2 Categories of efficacy estimands

Subject	Baseline covariate	Y(1)	Y(0)	δ	D	Y_{obs}
1	20	10	1	9	1	10
2	30	11	2	9	1	11
3	40	12	3	9	1	12
4	20	10	1	9	0	1
5	30	11	2	9	0	2
6	40	12	3	9	0	3

Table 15.3 Categories of efficacy estimands

Subject	Baseline covariate	Y(1)	Y(0)	δ	D	Y_{obs}
1	20	10	1	9	1	10
2	30	15	2	13	1	11
3	40	20	3	17	1	12
4	20	10	1	9	0	1
5	30	15	2	13	0	2
6	40	20	3	17	0	3

treated subject with a baseline value of 40 with the control subject with a baseline value of 20 leads to a biased estimate,

$$\hat{\delta} = 12 - 1 = 11 \neq 9 = \delta$$

This last point can be made a little more rigorous. The treatment comparison is fair if similar groups of subjects are compared in the two treatment arms. This will happen if the probability of observing a subject is independent of $\delta = Y(1) - Y(0)$ (i.e., a random sample from the population) or $D \perp\!\!\!\perp Y(1), Y(0)$.

Now consider Table 15.3 where δ is also correlated with the baseline covariate (i.e., sicker subjects have larger responses). Now, if the probability of observing a subject's response is proportional to the baseline, there will be a selection bias and both the individual treatment arm estimates and the treatment difference will be biased relative to the entire population.

In general, there are three ways to handle this situation:

1. Ignore the selection bias, and under the assumption it is small and not scientifically relevant, that is use available data only. This approach has the benefit of being transparent and easy to apply. The risk is that the estimate will be biased and misleading.
2. Use statistical methods that adjust for the bias to get an estimate that is unbiased for the entire population. If this is done well, the benefit is that you will have an

estimate for your original population that the experiment was designed for. The risk is that the statistical methods used to remove the bias are based on incorrect assumptions (i.e., you have incorrectly specified the conditional predictive distribution or the selection distribution of the last section). This can result in a failure to correct or even an increase in the bias.

3. Recognize that you no longer have an unbiased estimate of the entire population and alter your estimand to describe that fact (e.g., subjects with moderate to severe disease). This can be very useful but still requires assumptions and statistical methods to make sure the treatment comparison is unbiased in the subpopulation.

15.3 Clinical Trial Principles

Clinical colleagues will most likely not be conversant in estimands or current statistical practices for missing data. However, they like all of us have been exposed to a grab bag of 'principles' such as randomization, blinding, intention to treat (ITT), per protocol (PP) analyses, and simple imputation methods such as last observation carried forward (LOCF). They are all in some way related to estimands and missing data, but none of these alone define an estimand, in that they do not clearly state what the analysis is providing an estimate of. Let us review these principles and see how they relate back to missing data, estimands, and the framework described earlier.

15.3.1 Intention to Treat

No discussion of clinical trials is complete without the intention to treat (ITT) principle. The first author began working on clinical trials in the 1990s mentored by statisticians who began their careers in the 1960s and 1970s. They told tales of an earlier time when only subjects with complete data that were deemed appropriate for inclusion in the analysis were used. Unsurprisingly, trialists abused the privilege of having so much leeway to make their drugs look better and were thereafter forced to adhere to the intention to treat (ITT) principle by the regulatory community to avoid compromising the results of a trial.

What does the ITT principle really mean? This question is open for discussion. Hollis and Campbell (1999) reviewed all randomized controlled trials published in *BMJ*, *Lancet*, *JAMA*, and *NEJM* in 1997 and found inconsistent use and interpretation of the term. Informally, ITT is usually described as 'as randomized, as analyzed.' More detailed and thought out descriptions are fortunately available, and a working group of the biopharmaceutical section of the ASA in 1990 provided the following description:

> ... includes all randomised patients in the groups to which they were randomly assigned, regardless of their compliance with the entry criteria, regardless of the treatment they actually received, and regardless of the subsequent withdrawal from treatment or deviations from the protocol

(Lewis and Machin 1993). Lewis and Machin continue, '...the concept of ITT is clear in purpose and execution. These are parallel group studies of mortality and similar hard end-point in which medical or surgical intervention is compared to no intervention.' At the end of their article, they close with this defense of ITT, 'Anyone who follows these principles intelligently, and with a view to minimising bias, need not worry further about "intention to treat".' This is all wise advice, but how do you define bias? That was the question most were failing to ask at the time.

How does this relate to the theoretical framework? If data are all observed, then the ITT principle preserves randomization which ensures unbiased estimation. Unfortunately, it is not prescriptive for what to do if data are missing, that is what assumptions should we make about the conditional predictive distribution.

15.3.2 Per Protocol

A per protocol analysis includes only subjects who adhere to the protocol (which includes treatment compliance). The assumption is that this will provide an estimate of the maximum drug effect since only subjects who have complied with treatment and have had no other untoward events are analyzed. Furthermore, the hope was that the PP and ITT estimates would be similar, thus supporting the primary ITT analysis. The flaw in this argument is that the PP analysis does not respect the original randomization, thus allowing a potential selection bias, not necessarily in the direction that is expected. For example, consider a trial of an analgesic compared against a placebo control in subjects with chronic pain. If the analgesic works and is well tolerated, most treated subjects will remain in the PP population and provide a reasonable estimate of the drug effect. However, many placebo subjects will have inadequate pain control and require rescue medication or drop out of the trial and hence not be included in the PP population. The remaining placebo subjects will have lower pain scores than the entire placebo population, thus creating a bias that decreases the treatment effect. In other situations, the direction of the bias may be more difficult to predict.

15.3.3 Make the P Value Large

Since ITT is incomplete, it tells you to analyze everyone as randomized, but not what to do if they have missing data; some other principle is needed to help decide how to analyze a trial with missing data. In the absence of clear advice on how to account for missing data, the general rule of thumb given was to be 'conservative,' where the

definition of conservative is to make the p value for the primary efficacy endpoint as large as possible.

How does this relate to our framework? It completely ignores it. The point of the estimand framework is to find an unbiased estimate of a parameter of scientific and clinical interest. Merely taking an arbitrary estimator and making it more difficult to obtain statistical significance do not help to answer a scientific question.

15.3.4 Carry It Forward

One way to account for missing data is to 'fill in' or impute missing data for a subject using their observed values. There are a number of simple single imputation methods available for this purpose such as: last observation carried forward (LOCF), baseline observation carried forward (BOCF), and worst observation carried forward (WOCF). For example, if the endpoint of interest is average pain at week 12 and the subject withdrew right after their week 8 visit, you could carry forward that last observed value at week 8 and use it in lieu of the unobserved value at week 12. These methods all fall under the category of 'single imputation methods' which are not principled methods in that they ignore the additional uncertainty introduced by missing data. LOCF can underestimate the variance by using the same observation multiple times. In fact there is a large body of literature showing that LOCF estimates can be affected by many aspects of the data collection process including trends over time and can be have an inflated type I error rate (Mallinckrodt et al. 2011).

15.3.5 Ignore It

Little and Rubin's work on missing data (1987) became more widely appreciated in the clinical trial setting during the early 2000s, years after its original publication. Special attention was spent on the concept of data that is missing at random (MAR) and the associated concept of ignorability. As discussed earlier, there are a number of technical details associated with MAR and ignorability, based on the selection model factorization and the assumption of functionally distinct parameters. However, they were often boiled down to protocol language which stated the continuous longitudinal data in the study had values missing at random and therefore a mixed model repeated measures analysis (MMRM) would provide an unbiased treatment estimate. Again, the issue of bias is raised without much detail as what it really means, at least in protocol and statistical analysis plan text.

Statistical bias is the expected systematic difference between the estimate and the underlying parameter. The estimand defines that underlying parameter. So what is the implicit estimand for an MMRM analysis ignoring the selection distribution? Based on the results for ignorability, this implies we are assuming the data are MAR and the MMRM is correct. In that case, the model implicitly assumes that subjects with

incomplete trajectories over time are similar to subjects in the same treatment group, with similar covariates and similar observed responses who complete the study. In other words, it assumes that subjects remain on their randomized treatment after they withdraw from the study and therefore provide a type of efficacy estimand. This may not be desirable for regulators who want more of an effectiveness estimand assuming that subjects lose treatment benefit when they withdraw from the study.

15.4 Recent History

15.4.1 Panel on Handling Missing Data in Clinical Trials; National Research Council

In 2010, the 'Prevention and Treatment of Missing Data in Clinical Trials' was published by the National Research Council of the National Academies. This document is available to all online at no cost and was requested and funded by the FDA. Its release marked a major turning point in the demands of regulatory agencies for rigor in handling missing data in clinical trials. The document stated the importance of: estimands, designing trials to minimize missing data, following up subjects after treatment withdrawal, and included a technical overview of many methods of analysis.

It has 18 main recommendations. Number 1 was

> The trial protocol should explicitly define (a) the objectives(s) of the trial; (b) the associated primary outcome or outcomes; (c) how, when and on whom the outcome or outcomes will be measure; and (d) the measures of intervention effects, that is, the causal estimand of primary interest. These measures should be meaningful for all study participants, and estimable with minimal assumptions. Concerning the latter, the protocol should address the potential impact and treatment of missing data.

This document was treated as a beginning, not a final pronouncement on how to handle missing data; in fact, an article by FDA officials Robert O'Neill and Robert Temple describing the NAS document states '... the NAS panel's report on preventing and addressing missing data in clinical trials provides a roadmap, or perhaps a "problem list," for how we might proceed in the future.'

15.4.2 ICH E9 Addendum

In 2014, an addendum to ICH E9 was proposed that would include material on estimands and sensitivity analyses. A meeting was held in February 2015 to discuss the addendum. The results of this meeting were published in *Pharmaceutical Statistic* in 2017 by Phillips et al. The paper "Estimands: discussion points from the PSI

estimands and sensitivity expert group," included the following high-level summary:

> A clear message from the meeting was that estimands bridge the gap between study objectives and statistical methods. When defining estimands, an iterative process linking trial objectives, estimands, trial design, statistical and sensitivity analysis needs to be established. Each objective should have at least one distinct estimand, supported by sensitivity analyses.

The paper goes on to describe estimands in more detail and provides a useful framework for considering estimands in a clinical trial, but still fails to provide a concise definition.

15.5 Simulation Process and Estimands

In this section, we demonstrate a simulation study accounting for missing data. Where do estimands fit into this process? How is this different from merely considering missing data? The difference is subtle but important. Monte Carlo methods will be used to find the operating characteristics of the statistical analysis method under consideration. Monte Carlo methods in a nutshell entail the generation of many simulated trials which are: individually analyzed with the statistical method of interest, the results saved, and operating characteristics calculated by averaging over the simulations. In a single simulation, each subject will have a 'true' vector of observed responses y_{full} (with no missing observations) and an observed vector, y_{obs} (with possibly missing observations). For simplicity, assume a monotone missing data pattern (see Sect. 15.9), and then the two vectors, y_{full} and y_{obs} will be identical up to the time of dropout or significant protocol violation depending upon the study design and availability of retrieved dropout data. Note that there is a distinction between withdrawal from the study with no possibility of measuring a response at all and withdrawal from the protocol (e.g., low or no compliance, rescue medication use) where the response no longer reflects the protocol-specified treatment. After a subject's simulated dropout time or withdrawal from protocol, y_{obs} will either have a missing value or if the protocol allows, a value representing what would be seen if their response was still measured after they withdraw from treatment.

The key will be to generate y_{full} in a manner consistent with the estimand of interest. In other words, in the ideal trial we would analyze y_{full} to obtain an estimate for our estimand. Bias is then $E[y_{full} - y_{obs}]$ and likewise for other measures. During the simulation process, we can analyze y_{obs} and compare the results to analyzing y_{full} which does not require any imputations or other missing data methods and compare the results to find bias, power loss, etc.

There are two important points to this process. First, the 'true' state of nature is not known. We may assume that subjects who are intolerant to the active drug receive no benefit after discontinuing active drug. If so, do they immediately lose the drug effect, or is the loss gradual? A related question is does the drug modify the course of the disease or does it only provide symptomatic relief? An answer to these

questions requires scientific and clinical input; however, in many cases the opinions of the scientific experts will not be unanimous and simulations will need to be done under different scenarios.

The second point relates to the reason for dropout. Borrowing from the principal strata literature, we propose simulating latent causes for dropout. For example, a subject can be a 'non-tolerator' of active drug causing them to drop out, or they can have an unrelated adverse event causing dropout. However, in a real trial, we may not always be able to distinguish between the two types of adverse events. The simulations must also account for the effect of mis-identification of the true dropout category if that can affect a particular analysis method of interest.

15.6 Example: Simulation for Total Motor Score in Huntington's Disease

Our example is loosely based on redesigning a real phase II clinical trial in Huntington's disease. Huntington's disease is described as '.. an inherited disease that causes the progressive breakdown (degeneration) of nerve cells in the brain. Huntington's disease has a broad impact on a person's functional abilities and usually results in movement, thinking (cognitive) and psychiatric disorders.' (Mayo Clinic 2017).

The original study was designed to compare an experimental drug to placebo by measuring motor function in Huntington's disease subjects after 26 weeks of dosing; the primary endpoint was the change from baseline total motor score (TMS) compared to placebo. The total motor score is an assessment within the Unified Huntington Disease Rating Scale and ranges from 0 to 124; the larger the score, the worse the disease.

In the original study design, no estimand was specifically defined and the primary efficacy analysis was based on an MMRM approach assuming data were MAR. There was no plan to assess motor scores after treatment withdrawal or other significant protocol deviations; that is, there was no attempt to obtain retrieved dropout data. Without loss of generality, our example will use a two-arm parallel group design comparing placebo to a fictitious 20 mg dose. As in the original design, no retrieved dropout data will be made available for analysis.

15.7 Base Case

The primary endpoint is the change from baseline in the total motor score (TMS) assessment of the Unified Huntington Disease Rating Scale (UHDRS) after 26 weeks of treatment. The observed score ranges from 0 to 124. For our redesign, we will use historical placebo data measured at baseline and weeks 1, 2, 4, 8, 13, 19, and 26, as a basis for our simulated data, and add a treatment effect on to the baseline. Begin

by considering the 'base case,' that is the placebo TMS response in subjects who complete the trial and have no post-randomization deviations during the 6-month double-blind period.

In usual sample size consideration, the base case and the base case plus a treatment effect, δ, are the only scenario that is considered. Although longitudinal data like the TMS collected over time in this study may be analyzed using a mixed model repeated measure (MMRM) analysis, sample size calculations are often based on a cross-sectional model (e.g., ANCOVA or t-test) at the landmark visit of interest (6-month visit in this case) assuming no missing data, to get the number of complete subjects needed. The number of complete subjects is then inflated to account for dropout. For example, if 90 subjects are needed per arm for a t-test with sufficient power and a 10% dropout rate is expected, a sample size of 100 subjects will be reported to ensure 90 are available for analysis.

In order to respect the allowed range of TMS, absolute scores are simulated from a normal distribution and then truncated to fall in the allowable range of 0–124. The scores could have also been truncated further to reflect inclusion criteria for baseline TMS, however, that was not done for the simulation results that follow. This is less complicated than making adjustments to change from baseline scores. In addition, this also allows the flexibility of having simulated subjects with no post-baseline data if so desired.

The mean values for the base case are displayed in Fig. 15.1. The figure displays a stylized fact for placebo response in Huntington's disease where the response improves by 2–4 points and lasts approximately 3 months (de Yebenes et al. 2011). The correlations over time are modeled by a combination of a random subject effect and an AR(1) process for residual correlation. This structure can have a strong correlation among all visits but still allow for a decreasing correlation as the time interval

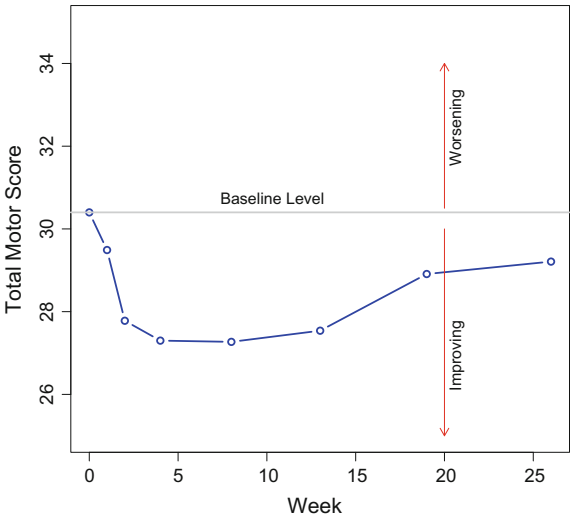


Fig. 15.1 Base case: mean profile

Table 15.4 Within subject correlation matrix

Week	0	1	2	4	8	13	19	26
0	1.00	0.91	0.89	0.88	0.87	0.87	0.87	0.87
1	0.91	1.00	0.91	0.89	0.88	0.87	0.87	0.87
2	0.89	0.91	1.00	0.91	0.89	0.88	0.87	0.87
4	0.88	0.89	0.91	1.00	0.91	0.89	0.88	0.87
8	0.87	0.88	0.89	0.91	1.00	0.91	0.89	0.88
13	0.87	0.87	0.88	0.89	0.91	1.00	0.91	0.89
19	0.87	0.87	0.87	0.88	0.89	0.91	1.00	0.91
26	0.87	0.87	0.87	0.87	0.88	0.89	0.91	1.00

between visits grows longer, while only requiring the specification of three parameters. The mixed model representation is

$$y_{ij} = \mu_{ij} + b_i + e_i j, \tag{15.3}$$

where μ_{ij} is the fixed mean effect for patient i at visit j and b_i is a random subject-specific term with a $N(0, \sigma_b^2)$ distribution. The error terms model both the correlation between visits and the measurement error. For example, if there were three visits,

$$Var(e) = \sigma_e^2 \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}. \tag{15.4}$$

In general, $Var(e_{ij}) = \sigma_{ij} = \sigma_e^2 \rho^{|i-j|}$. For this example, we have chosen $\sigma_b^2 = 164$, $\sigma_e^2 = 24$ and $\rho = 0.33$. The resulting correlation matrix is shown in Table 15.4.

The careful reader may have noticed that technically the AR(1) structure assumes that the time intervals between visits are all equal. However, based on empirical data the structure fits well enough, and this may be true since the correlation structure is mostly determined by the random subject effect.

The treatment effect is unknown and needs to be hypothesized. We generate different treatment effect scenarios by defining the treatment effect with a two-step process:

1. Define a maximum treatment effect δ_{max} that can be varied from a null effect $\delta_{max} = 0$ to any desired magnitude.
2. A vector of fractional effects $f = \{f_0, f_1, f_2, f_4, f_8, f_{13}, f_{19}, f_{26}\}$. This vector defines the shape of the treatment effect over time and is held fixed.

For example, consider the following two possible treatment effects:

1. Linear effect $f = \{0.00, 0.04, 0.08, 0.15, 0.31, 0.50, 0.73, 1.00\}$.
2. Linear followed by flat effect $f = \{0.00, 0.20, 0.40, 0.60, 0.80, 1.0, 1.0, 1.00\}$.

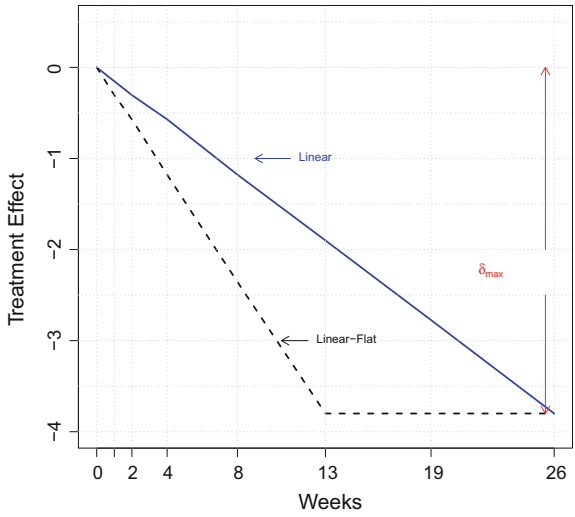


Fig. 15.2 Treatment effect by time

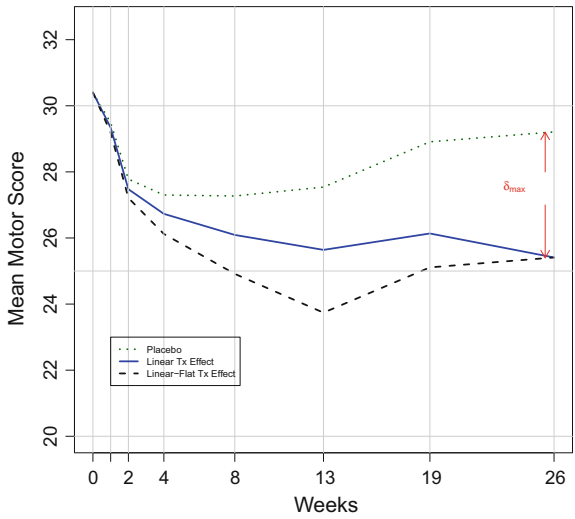


Fig. 15.3 Means by time

Thus, the treatment effect at time t is $f_t \times \delta_{max}$. Note that δ_{max} will be a negative number since lower total motor scores are associated with higher functioning. It may be best to think of δ_{max} as the maximum absolute treatment difference. These are displayed in Figure 15.2 where δ_{max} is taken to be -3.8.

Figure 15.3 displays motor scores by time for the base case placebo group and for treatment groups by superimposing both the linear time course and linear to flat time course, with a δ_{max} of -3.8 on top of the base case. Only, one of the two time courses would be chosen for a single simulation.

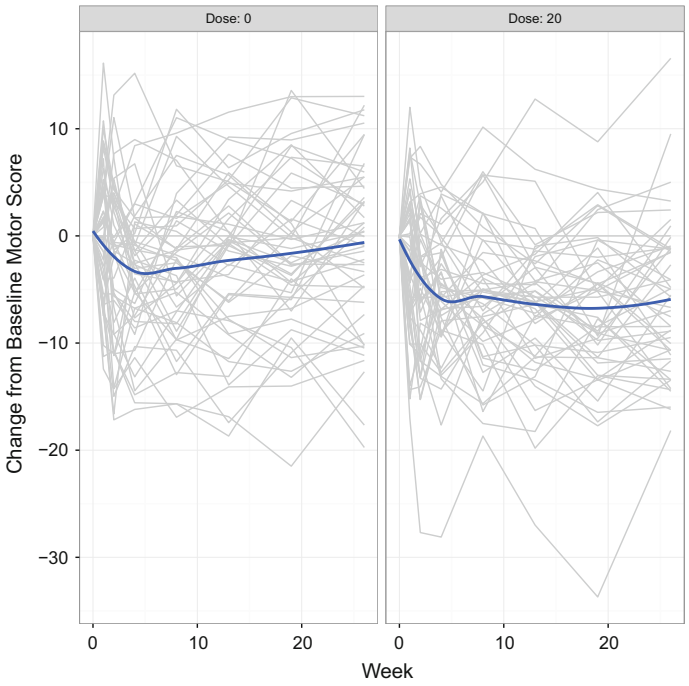


Fig. 15.4 Simulated change from baseline motor score trajectories

For the simulations, observed motor scores were first simulated from a multivariate normal distribution using the base case, a linear treatment trend and a δ_{max} of -3.8 with random correlated error as specified in the compound symmetry plus AR(1) covariance structure described earlier. The simulated data are then truncated to [0, 124] to match the range of TMS scores. Lastly, change from baseline scores is calculated using the simulated truncated TMS scores. As an example, simulated change from baseline scores for 50 placebo and 50 active subjects is displayed in Fig. 15.4.

The results from 10,000 simulated datasets with no missing data are displayed in Table 15.5. The table has three columns, with results from an MMRM analysis with all post-baseline change scores and baseline as a covariate, an ANCOVA at month 6 using the baseline motor score as a covariate and a simple t-test of the change from baseline data at month 6. The 'Delta' row displays the average estimated treatment difference in the change from baseline averaged over the 10,000 simulations, while the 'SD' row displays the standard deviation of the 10,000 estimated treatment differences. The last row displays the proportion of simulations that led to rejection of the null hypothesis, $\delta = 0$, using an $\alpha = 0.05$ level two-sided test.

Note that all three analyses which are expected to be unbiased for a δ of -3.8 had an average δ of -3.72. This apparent bias is due to the small number of simulated patients who had raw scores of less than 0 at month 6 that were then truncated as

Table 15.5 Simulation of completer data B = 10,000

	MMRM	ANCOVA	t-test
Delta	−3.72	−3.72	−3.72
SD	1.31	1.31	1.36
Power	0.83	0.80	0.77

described above. The resulting truncated normal distribution does of course have a different mean than the underlying normal distribution, but specifying those moments in advance is difficult. To account for this in the simulations to follow, we calculate bias relative to an analysis of Y_{full} instead of the theoretical δ. This approach can be used for any simulation regardless of how complex as long as enough simulations are produced to minimize Monte Carlo error. It is also noteworthy that the MMRM analysis has 3% greater power than the ANCOVA using only landmark data and 6% greater power than the t-test which does not incorporate the baseline score as a covariate.

15.8 Estimands

For this simulation study, we consider two extreme cases for the estimands to better highlight the differences between effectiveness and efficacy estimands: one, an effectiveness estimand, where we seek an estimate for a population that discounts the treatment effect for subjects who do not comply with the protocol and the second an efficacy estimand, meaning we want to estimate the drug effect if taken as directed (assuming a subject can tolerate the drug).

1. Effectiveness
   - Population: Huntington's disease subjects as defined by the inclusion and exclusion criteria of the study.
   - Endpoint: Change from baseline in total motor score at 6 months.
   - Measure of Intervention Effect: Difference in mean change from baseline in total motor score at 6 months in the placebo and 20 mg treated groups *regardless of tolerability and compliance*.
2. Efficacy
   - Population: Huntington's disease subjects as defined by the inclusion and exclusion criteria of the study.
   - Endpoint: Change from baseline in total motor score at 6 months.
   - Measure of Intervention Effect: Difference in mean change from baseline in total motor score at 6 months in the placebo and 20 mg treated groups *if all subjects had adhered to the study*.

The effectiveness estimand is in the spirit of an ITT estimate. It is concerned with the effect of the 'treatment policy', that is what is the effect of prescribing the active drug. If subjects have less than perfect compliance, withdraw from treatment, or take rescue medication, it is still possible to measure their outcomes and use this, retrieved dropout, data to get an unbiased estimate of the effectiveness estimand. In contrast, the efficacy estimand is interested in the hypothetical scenario that everyone adheres to the treatment. If a subject stops taking drug or initiates a rescue medication, then their observed scores cannot be used as is and are set to missing.

Both estimands can in some circumstances yield paradoxical results. A pure effectiveness estimand could attribute a positive effect to placebo because the majority of placebo patients began taking a rescue medication, thus providing a comparison of new drug versus rescue medication instead of placebo. The pure efficacy estimand can conversely overstate the effect of the new drug by attributing a positive effect to subjects who cannot tolerate or otherwise take the medication for any sustained length of time. Therefore, it may make more sense to modify these two extreme estimands somewhat so that placebo subjects are not attributed a positive effect by switching to rescue medication and actively treated subjects are not attributed a positive effect for taking a drug they cannot tolerate. This is the aim of estimand 6 in Sect. 15.1. In the case of an efficacy estimand, it may also make sense to limit the estimand to the population of those who can tolerate the drug. There are analytical methods that attempt to do this; however, the best method is to use an enriched study design that includes a run-in period where all subjects are exposed to study drug and only those who tolerate the drug move forward to the randomized portion of the trial. Thus, different estimands may require different study designs making the comparison of estimands and statistical methods more complicated.

15.9 Dropout Reasons and Patterns

We assume there are five categories of subjects:

1. **Completers**, who will complete the study regardless of treatment assignment, and provide all protocol-specified data.
2. **Loss to Follow-up** subjects, who will eventually leave the study for reasons unrelated to treatment and disease. This category will also include subjects who are unwilling to participate and other reasons unrelated to disease and drug.
3. **Non-compliers** who take anywhere from 0 to 80% of assigned drug.
4. Subjects with **related adverse events** whose adverse event is related to study drug and leads to drug discontinuation. If the protocol design includes follow-up on patients who remain on study but off treatment (or protocol in general) data may be available for some of these subjects. These subjects can only be identified in the observed data if they were randomized to active drug.
5. Subjects with **unrelated adverse events** that are not caused by study drug (Table 15.6).

Table 15.6 Dropout reasons

Category	Active treatment (%)	Placebo (%)
Completer	66	86
Loss To FUP	4	4
Non-complier	3	3
Related AE	20	0
Unrelated AE	7	7

Table 15.7 Discrete dropout time distribution

Week	0	1	2	4	8	13	19	26
Completer	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Loss to FU	2.0	7.5	7.5	7.5	12.5	16.0	22.0	25.0
Non-complier	0.0	0.0	20.0	40.0	20.0	10.0	10.0	0.0
Related AE	0.0	0.0	20.0	40.0	20.0	10.0	10.0	0.0
AE	12.5	12.5	12.5	12.5	12.5	12.5	12.5	12.5

Table 15.8 Monotone and non-monotone missing data

Week	Baseline	1	2	4	8	13	19	26
Monotone	40	39	38	37	35	?	?	?
Non-monotone	40	36	38	?	35	?	?	?

? marks indicate missing data

Once a subject is randomly assigned to a non-completer group, a dropout time also needs to be randomly chosen. This will be done via a discrete distribution on the visit times. The distribution chosen is displayed in Table 15.7

For simplicity, only monotone dropout patterns will be generated; that is, a subject will miss their dropout visit and all subsequent visits. A non-monotone dropout pattern would have intermittent missing visits. Table 15.8 illustrates the difference between monotone and non-monotone patterns.

Another simplifying assumption for these simulations is that dropout time is only dependent upon the reason for dropout. A simulation could alter this to account for baseline severity or any other covariates of interest.

15.9.1 Full and Observed Data

The simulations will create a 'full' data vector for each subject and an observed data vector with missing values. Both the simulations and the statistical analyses require assumptions about the true full data, which is unknown to us. Therefore, it may be

Table 15.9 Effectiveness simulation scenario

	Category	Full data		Observed data	
		Active treatment	Placebo	Active treatment	Placebo
Before drop	Completer	$BC + \delta$	BC	$BC + \delta$	BC
	Loss To FUP	$BC + \delta$	BC	$BC + \delta$	BC
	Non-complier	BC	BC	BC	BC
	Related AE	$BC + \delta$	BC	BC	BC
	Unrelated AE	$BC + \delta$	BC	$BC + \delta$	BC
Post-drop	Completer	–	–	–	–
	Loss To FUP	BC	BC	Missing	Missing
	Non-complier	BC	BC	Missing	Missing
	Related AE	BC	BC	Missing	Missing
	Unrelated AE	BC	BC	Missing	Missing

wise to test the analysis assumptions against different possible 'true' cases to see how much bias and type I error inflation can occur.

Due to the original project that initiated this study, we consider a two-arm parallel group trial with no retrieved dropout data and start by defining the mean effect for y_{full} and y_{obs}. The correlation structure will be taken from base case and applied to all groups. This may not be true, but we do not have sufficient historical information to define multiple correlation structures.

Starting with the effectiveness estimand, we assume that placebo subjects will have on average the mean structure defined for the base case, both pre- and post-dropout. Treated subjects will all have the same treatment effect prior to dropout, except the non-compliers who we believe do not receive enough study drug to show a treatment effect. Finally, all treated subjects are assumed to return to the base case scenario after dropout by the next visit. This is reasonable for a drug that works symptomatically so that treatment effects are observed in addition to the normal progression of the disease. If the disease progresses quickly and the drug modifies the course of treatment, then a gradual decline from their treated state would make more sense. This effectiveness scenario is displayed in Table 15.9.

The efficacy simulations will work in a similar manner, except now all post-dropout data will have the same mean effect as the pre-dropout data. This is summarized in Table 15.10.

15.9.2 Analysis Methods

All analyses are applied to change from baseline TMS, for the HD disease trial described earlier. In all cases, data are monotonically missing after dropout,

Table 15.10 Efficacy simulation scenario

	Category	Full data		Observed data	
		Active treatment	Placebo	Active treatment	Placebo
Before drop	Completer	$BC + \delta$	BC	$BC + \delta$	BC
	Loss To FUP	$BC + \delta$	BC	$BC + \delta$	BC
	Non-complier	BC	BC	BC	BC
	Related AE	$BC + \delta$	BC	BC	BC
	Unrelated AE	$BC + \delta$	BC	$BC + \delta$	BC
Post-drop	Completer	–	–	–	–
	Loss To FUP	$BC + \delta$	BC	Missing	Missing
	Non-complier	$BC + \delta$	BC	Missing	Missing
	Related AE	$BC + \delta$	BC	Missing	Missing
	Unrelated AE	$BC + \delta$	BC	Missing	Missing

non-compliance, adverse event, or loss to follow-up. Only the analysis methods change in the simulation results that follow. The analysis methods are:

1. Multiple Imputation (**MI**) is a very simple application of an approximate Bayesian bootstrap (ABB). The algorithm has two steps: First identify a set of 'donor' data and take a bootstrap sample. The bootstrap sample is then bootstrapped again to impute missing values.
   - Effectiveness: Complete placebo subject data at week 26 were used for the donor data.
   - Efficacy: Complete data from subjects in the same treatment arm at week 26 are used for the donor data.
2. **MMRM** is a linear mixed model repeated measure analysis using all available data. The model includes treatment and baseline TMS score as a fixed effect and a random subject effect. The model also includes an autoregressive structure for residual correlation.
3. **ANCOVA** is a linear model using available week 26 data with terms for treatment and baseline TMS score.
4. **t-test** compares the mean values of the available week 26 data. This is the same as an ANOVA analysis.
5. **ANCOVA LOCF** is an ANCOVA using available and last observation carried forward change from baseline data at week 26.
6. **ANCOVA-Full** is an ANCOVA using y_{full} and represents the 'ideal' analysis.

15.9.3 Simulation Results

Tables 15.11 and 15.12 display the results for a 50-subject per arm trial for the effectiveness and the efficacy estimands under the linear treatment effect, with 10,000 Monte Carlo simulations. The sample size of 50 was of interest to us, and we wished to see what the operating characteristics would be for this size of trial. The first panel in each table where $\delta_{Comp} = 0$ displays the results under the null hypothesis of no treatment effect. For both effectiveness and efficacy, all methods control the type I error at approximately the 0.05 level, with the possible exception of a small amount of inflation for the MMRM analysis.

In Table 15.11, we see the only unbiased method for the effectiveness estimand is the MI approach as demonstrated by the close agreement of the average simulated Delta estimates for the MI and ANCOVA-Full columns. The average Deltas are similar for MMRM, ANCOVA, and the t-test which all overstate the treatment effect because they all provide an estimate of efficacy, not effectiveness, as demonstrated by their close agreement with δ_{Comp} for each panel of the table, where δ_{Comp} is the simulated treatment difference in the completers prior to truncating the distribution. The full drug effect is slightly less than δ_{Comp}. How much less? We can see that in Table 15.12 where the reported Deltas for ANCOVA-Full show the full treatment effect. Averaging the ratio of Delta for ANCOVA-Full divided by δ_{Comp} over the efficacy simulation scenarios in Table 15.12 results in a Delta that is approximately 97% of δ_{Comp}. The same ratio calculated over the effectiveness simulation scenarios in Table 15.11 results in a Delta that is approximately 64% of δ_{Comp}.

For the efficacy estimand in Table 15.12, all of the analysis methods except for LOCF are unbiased, with MMRM clearly having the advantage in power. As a matter of fact, LOCF is also biased for the effectiveness estimand. LOCF estimates the treatment effect at the last observed visit. Since there is downward treatment effect over time, every treated subject with missing data uses an observation prior to week 26 with a less than maximal treatment effect. Therefore, LOCF targets an estimand somewhere between effectiveness and efficacy.

15.9.4 Sample Size

The simulations presented explore the operating characteristics of different analysis methods for two different estimands with a fixed sample size of 50 per arm. This is a common situation when beginning to design a study where historical knowledge and logistical considerations often suggest a reasonable sample size. However, there is also usually a treatment effect size that is of concern. The magnitude of this effect size can be determined by either what is clinically meaningful or sufficiently large compared to a competitor product. Therefore, we must also explore whether a study needs to be increased in size or if it can be decreased. In usual practice, a treatment difference, δ or effect size, $ES = \delta/\sigma$, where σ is the standard deviation in a single

Table 15.11 Effectiveness estimand: N/arm = 50, linear treatment effect

$\delta_{Comp} = 0$	MI	MMRM	ANCOVA	t-test	ANCOVA LOCF	ANCOVA-Full
Delta	0.01	0.01	0.01	0.00	0.01	0.02
SD	1.17	1.60	1.68	1.73	1.34	1.34
Power	0.01	0.07	0.05	0.05	0.05	0.05
$\delta_{Comp} = -1.9$						
Delta	−1.22	−1.85	−1.85	−1.86	−1.34	−1.21
SD	1.18	1.60	1.68	1.73	1.33	1.35
Power	0.06	0.25	0.20	0.19	0.18	0.15
$\delta_{Comp} = -3.8$						
Delta	−2.44	−3.69	−3.70	−3.71	−2.69	−2.43
SD	1.19	1.59	1.67	1.73	1.34	1.36
Power	0.27	0.69	0.60	0.58	0.53	0.43
$\delta_{Comp} = -5.0$						
Delta	−3.21	−4.85	−4.86	−4.87	−3.53	−3.20
SD	1.21	1.59	1.67	1.73	1.35	1.38
Power	0.49	0.89	0.83	0.80	0.75	0.64
$\delta_{Comp} = -6.0$						
Delta	−3.84	−5.80	−5.82	−5.83	−4.23	−3.83
SD	1.23	1.59	1.67	1.73	1.35	1.39
Power	0.67	0.96	0.93	0.92	0.88	0.78
$\delta_{Comp} = -7.6$						
Delta	−4.84	−7.32	−7.34	−7.34	−5.34	−4.83
SD	1.26	1.59	1.67	1.73	1.37	1.42
Power	0.87	1.00	0.99	0.99	0.97	0.92
$\delta_{Comp} = -10$						
Delta	−6.32	−9.55	−9.58	−9.58	−6.97	−6.31
SD	1.33	1.59	1.66	1.74	1.41	1.48
Power	0.98	1.00	1.00	1.00	1.00	0.99

arm, is held fixed, and the sample size n is varied. The estimand approach complicates this; what δ should we use for sample sizing?

Using the estimand approach, there are multiple δs, but the δ capturing the treatment difference in the base case is the one most closely linked to the underlying PK/PD of the drug. Effectiveness and various other measures of treatment difference will tend to have smaller magnitude and are in some sense arbitrary; that, is we decide if we are describing a population on a continuum from everyone has the full drug effect to no one has the full drug effect.

For our case, there was an interest of seeing a δ of -4.5. Unfortunately, this request was not specified in terms of estimands; is this an effectiveness or efficacy scenario?

Table 15.12 Efficacy estimand: N/arm=50, linear treatment effect

$\delta_{Comp} = 0$	MI	MMRM	ANCOVA	t-test	ANCOVA LOCF	ANCOVA-Full
Delta	0.02	0.00	0.00	0.01	−0.00	−0.01
SD	1.72	1.57	1.64	1.68	1.32	1.33
Power	0.06	0.06	0.05	0.05	0.05	0.05
$\delta_{Comp} = -1.9$						
Delta	−1.85	−1.86	−1.86	−1.85	−1.36	−1.87
SD	1.72	1.56	1.64	1.68	1.32	1.33
Power	0.19	0.25	0.20	0.19	0.18	0.29
$\delta_{Comp} = -3.8$						
Delta	−3.70	−3.70	−3.71	−3.70	−2.71	−3.72
SD	1.72	1.56	1.64	1.68	1.32	1.33
Power	0.55	0.69	0.61	0.58	0.53	0.79
$\delta_{Comp} = -5.0$						
Delta	−4.86	−4.86	−4.87	−4.86	−3.55	−4.88
SD	1.71	1.56	1.63	1.68	1.33	1.32
Power	0.77	0.89	0.83	0.81	0.75	0.95
$\delta_{Comp} = -6.0$						
Delta	−5.81	−5.81	−5.83	−5.81	−4.25	−5.84
SD	1.71	1.56	1.63	1.68	1.34	1.32
Power	0.90	0.97	0.94	0.92	0.88	0.99
$\delta_{Comp} = -7.6$						
Delta	−7.33	−7.33	−7.35	−7.33	−5.36	−7.36
SD	1.71	1.55	1.63	1.69	1.36	1.32
Power	0.98	1.00	0.99	0.99	0.98	1.00
$\delta_{Comp} = -10$						
Delta	−9.57	−9.56	−9.60	−9.57	−6.99	−9.60
SD	1.71	1.55	1.62	1.69	1.39	1.32
Power	1.00	1.00	1.00	1.00	1.00	1.00

If we want to simulate data with a treatment difference of −4.5 under an efficacy scenario, we know from our earlier simulations that mean difference in the truncated normal distributions describing the treatment groups is about 97% of the δ_{Comp} used in the simulation program; therefore, we used the same programs as before by setting $\delta_{Comp} = -4.62$ to obtain an efficacy difference of −4.5. Using similar logic and the fact that the mean difference in the truncated normal distributions describing the treatment groups for effectiveness is about 64% of the δ_{Comp} used in the simulation program, a δ_{Comp} −7.06 is needed to get an effectiveness treatment difference of −4.5. Immediately, we see that for a given treatment difference different estimands can require a markedly greater drug effect.

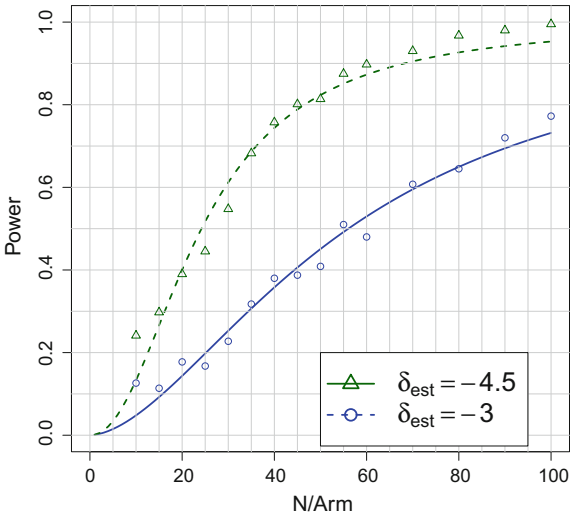


Fig. 15.5 Effectiveness—power curve

Figure 15.5 displays sample size versus power for the effectiveness scenario. Although an underlying treatment difference of −7.06 is needed to get and effectiveness estimand of −4.5, we also display the results when using an underlying treatment difference of −4.62. As shown in the figure legend, the resulting effectiveness estimands are −4.5 as expected and −3.0. Simulations were run for various sample sizes, and then a sigmoidal curve was fitted to help smooth out the results and enforce increasing power with increasing sample size. A sample size of 45–50 is sufficient for 80% power with an effectiveness estimand of −4.5 and underlying treatment effect of −7.06. The smaller treatment effect requires a sample size of greater than 100 (not shown).

Figure 15.6 is similar and displays sample size versus power for the efficacy scenario. Again, a sample size of approximately 45 per arm is sufficient for 80% power if the efficacy estimand difference is −4.5. Using an underlying drug effect of −7.06 results in an efficacy difference of −6.86 (again due to truncation effect) and has sufficient power with approximately 20 per arm.

15.9.5 Approximate Calculations

It is dangerous to rely on computer simulations without first having a rough idea of what the results should look like. In these simulation scenarios, we can make reasonable predictions of the simulation results, with a minimal amount of computations.

The power shown in Tables 15.11 and 15.12 is largely driven by the treatment effect and hence the effect size. The PMM approach used in setting up the simulation scenarios is particularly amenable to calculating a rough estimate of the effect size.

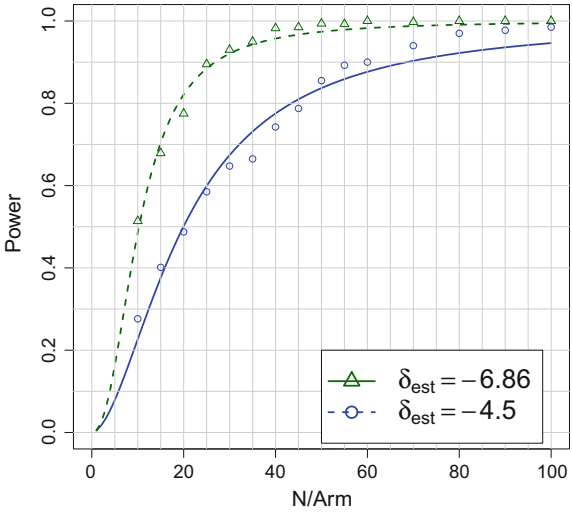


Fig. 15.6 Efficacy—power curve

For example, start with the efficacy estimand in Table 15.11 and the power in the last column using the complete data. The simulations assume a 66% completion rate for active subjects and 86% for placebo subjects (a 76% overall completion rate). All placebo subjects are assumed to follow the base case regardless of whether they drop out or not. Actively treated subjects, on the other hand, can be grouped as completers with a full δ_{Comp} effect and non-completers who revert to the base case. Therefore, the expected δ under this PMM approach is:

$$E(\delta) = 0.66 \times \delta_{Comp} + 0.34 \times 0 = 0.66 \times \delta_{Comp}.$$

The expected δ's are 0, −1.25, −2.51, −3.3, −3.96, −5.02, −6.6 which are all slightly smaller in magnitude than the average simulated δ, 0.02, −1.21, −2.43, −3.20, −3.83, −4.83, and −6.31 from the last column of Table 15.11, due to the truncation effect discussed earlier. The simulations also assumed a standard deviation at each time point of 13.7 and a correlation of 0.87 between baseline and week 26 data. Therefore, the standard deviation of the change from baseline score is:

$$SD(Y_{26} - Y_0) = \sqrt{2 \times 13.7^2 - 2 \times 0.87 \times 13.7^2} = 6.99,$$

not accounting for any truncation effect.

Given the expected differences, standard deviation, and sample size, the power of a hypothesis test can be calculated using standard methods. Using the R function power.t.test, and the full sample size of 50 per arm, we get: 0.05, 0.15, 0.43, 0.65, 0.80, 0.94, and 1.00, very nearly identical to the simulated power results in the last column for ANCOVA-Full. The MI results are also of interest because they are approximately unbiased for the true estimand, as seen by very close agreement between the average

estimated δ's in the simulation results. However, for these calculations we want to account for the loss in power due to missing data; therefore, we repeat the standard t-test power calculations using a sample size of $0.76 \times 50 = 38$, because only 76% of subjects are expected to complete the study. Repeating the calculations above with power.t.test with the same effect size and a new sample size of 38, we obtain power calculations of: 0.05, 0.13, 0.34, 0.53, 0.68, 0.87, 0.98.

A back of the envelope calculations can be used for the power curves as well. A large sample approximation for two sample tests tells us that

$$N_{arm} \propto \frac{1}{ES^2} \propto \frac{1}{\delta^2}.$$

In Fig. 15.6, 45/arm provides 80% power for a δ of -4.5. The sample size for $\delta = -6.86$ is about 20. Using the large sample approximation, we get

$$\frac{45}{N_{-6.86}} = \frac{-6.86^2}{-4.5^2},$$

or 19.4. This type of calculation can be used to spot check results for consistency and reduce calculations.

15.10 Discussion

Estimands clarify the scientific objectives of inference. These objectives are of interest to everyone on a clinical trial and provide the common language to unite statistical, medical, scientific, and regulatory experts. Statisticians can have a greater impact and be of more use to their clinical trial colleagues if they speak in the common language of scientific objectives instead of introducing statistical jargon to the conversation.

Estimands are not new to statisticians but rather a reminder of the importance of keeping study objectives and statistical analyses aligned. Furthermore, they keep the focus on parameters that have a clinical and scientific meaning, putting estimation first and hypothesis testing second, since there is little value in performing a statistically accurate hypothesis test on an estimate with little scientific value.

Another lesson learned long ago in the statistics community is that good design is essential and often more important than the analysis. For example, in the Huntington's disease simulation, if we truly wanted to design a trial for an effectiveness estimand, every effort should have been made to obtain TMS data even after a subject withdrew from treatment or had a protocol violations. This will not be possible for all subjects, but there should be enough information gained from the retrieved dropout data to help inform the conditional predictive distribution for those who could not be followed. If an efficacy estimand was of primary importance, then an enriched design of some sort would have been more appropriate.

The takeaway message is to design the trial you need from the start, with the estimand you need, and if you can speak in terms of scientific objectives, there are clinicians and scientists who can help.

References

Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*. *Statistics in practice*. Wiley.

de Yebenes et al (2011). Pridopidine for the treatment of motor function in patients with Huntington's disease (MermailHD): A phase 3, randomised, double-blind, placebo-controlled trial. *Lancet Neurol, 2011*(10), 1049–57.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Analytical Methods for Social Research. Cambridge University Press.

Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ*, *319*(7211), 670–674.

Lewis, J. A., & Machin, D. (1993). Intention to treat—Who should use ITT? *British Journal of Cancer*, *68*(4), 647–650.

Little, R. J., & Rubin, D. B. (1987). *Statistical Analysis With Missing Data*. Wiley.

Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health*, *21*(1), 121–145.

MacKay, R. J., & Oldford, R. W. (2000). Scientific method, statistical method and the speed of light. *Statistical Science*, 15(3), 254–278.

Mallinckrodt, C. H., Clark, W. S., & David, S. R. (2011). 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.

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.

Mayo Foundation for Medical Education and Research. Huntington's disease. Retrived June, 2017 from http://www.mayoclinic.org/diseases-conditions/huntingtons-disease/basics/definition/con-20030685.

National Research Council. (2010). *The prevention and treatment of missing data in clinical trials*.

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. OUP Oxford: Oxford science publications.

Phillips, A., Abellan-Andres, J., Soren, A., Bretz, F., Fletcher, C., France, L., Garrett, A., Harris, R., Kjaer, M., Keene, O., Morgan, D., O'Kelly, M., Roger, J. (2017). Estimands: Discussion points from the PSI estimands and sensitivity expert group. *Pharmaceutical Statistics*, 16(1), 6–11. PST. 1745.

Chapter 16
Development of Prognostic Biomarker Signatures for Survival Using High-Dimensional Data

Richard Simon

The heterogeneity of prognoses of patients with apparently the same type of cancer (i.e., same primary site and tumor histology) has long been recognized. This recognition has led to many attempts to develop prognostic models based on clinical and pathological factors. The development of genome-wide assays such as microarrays and next-generation sequencing has provided increased information for development of prognostic models. Such assays, however, provide information on the expression level and mutation status of all 20,000 plus human genes. For the information to be utilized for prognostic modeling, methods must be able to handle data where the number (p) of candidate predictors is vastly greater than the number of cases. This has stimulated the development of such methodology by the statistics and machine learning communities. In this chapter, I will review some of these methods for developing and evaluating prognostic models based on survival or other time-to-event endpoints.

16.1 Medical Utility of Prognostic Modeling

Prognostic models are generally based on baseline measurements and provide information about the likely long-term outcome of patients either untreated or with standard treatment. Too often the effect of treatment is ignored in the planning of the modeling effort because of lack of clarity on the intended use of the model. This may result in inappropriate selection of cases for model development and lack of utility of the resulting model. Subramanian and Simon (2010) reviewed the literature of prognostic models for patients with operable non-small-cell lung cancer. They found that lack of clarity on intended use of the model was very common. Although

R. Simon (✉)
R Simon Consulting, Rockville, MD 20850, USA
e-mail: rmaceysimon@gmail.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_16

the oncology literature is replete with publications on prognostic modeling, very few of these are used in clinical practice. Prognostic models can be useful if they inform physicians and patients with regard to treatment decisions. Unfortunately, most prognostic factor studies are conducted using a convenience sample of patients whose tissues are available. Often these patients are too heterogeneous with regard to treatment, stage, and standard prognostic factors to support therapeutically relevant conclusions.

Perhaps, the most successful example of prognostic modeling using the new assays was the development of the OncoType DX recurrence score for patients with stage I breast cancer (Paik et al. 2006). The prospectively defined intended use of the model was to determine which patients with stage I, and hormonal receptor-positive breast cancer had such good prognosis with only hormonal therapy that they do not require chemotherapy. The intended use determined that the cases selected for study all had stage I, hormonal receptor-positive tumors and did not receive chemotherapy. The intended use also determined the analysis of the data: merely finding a cut-point for the outcome of the model that identified a subset of patients with such good outcome that they might opt for not receiving chemotherapy. The 21 genes used in the model were determined based on previous studies, and no aspect of the analysis involved evaluating the relative value of individual genes.

16.2 Survival Risk Prediction Models

16.2.1 Penalized PH Modeling

The most commonly used method is penalized proportional hazards regression. The proportional hazards model is

$$\log\left\{\frac{h(t,X)}{h_0(t)}\right\} = \beta' X$$

where h(t, X) is the hazard function at time t for a patient with covariate vector X and $h_0(t)$ is the baseline hazard function. Usually, the PH model is fit by finding the vector of regression coefficients that maximize the partial log-likelihood function L. When the number of features is larger than the number of cases, however, maximization of L would result in extreme over-fitting of the data and would provide a model that predicts very poorly. As an extension of Tibshirani's Lasso method to survival data (Tibshirani 1997), maximization of L is replaced with maximization of a penalized log-likelihood:

$$\hat{\beta} = \text{argmax}_{\beta}\{\log(L(\beta) - \lambda||\beta||_1\} \tag{16.1}$$

where $\|\beta\|_1$ denotes the so-called L_1 norm of β; that is, the sum of absolute values of the components of β. For any nonzero λ value, some of the regression coefficients will be shrunken to zero. The larger the shrinkage parameter, the fewer features that remain in the model. Consequently, this method provides for both feature selection and regression parameter estimation. The penalty parameter λ is generally determined by cross-validation of the full dataset to maximize the predictive log-likelihood

$$\sum_{i=1}^{n} \delta_i \log\left(L_i(\hat{\beta}^{(-i)})\right) \tag{16.2}$$

where L_i denotes the partial likelihood factor for survival of the *i*th case evaluated using the vector of regression coefficients determined when the *i*th case was omitted from the training set. There is a term in the sum for each death event as δ_i is the censoring indicator.

An alternative approach, similar to ridge regression, is based on determining the regression coefficient vector using (16.1) but with the L2 norm (square root of sum of squares of components of the beta vector) replacing the L1 norm (Hastie and Tibshirani 2004; Van Houwelingen et al. 2006). Using the L2 norm, however, does not shrink regression coefficients to zero; positive coefficients may be shrunken to negative values. Consequently, it does not reduce the number of features in the model. For this reason, L2 shrinkage is often combined with an initial step for reducing the number of features. This may be done by selecting only the features with the greatest univariate association with survival. A third method for PH modeling is to use as predictors the first few "supervised principle components" of the features. The supervised principle components are the regular principle components when the calculation is restricted to features which have a strong univariate association with survival (Bair and Tibshirani 2004).

Some investigators have used the first few regular principle components of the features as predictors in a PH regression model. This often fails to provide good survival risk models because the principle components are computed to maximize variability but not for good correlation with survival. The method of partial least squares, originally developed for quantitative uncensored outcomes, has been adapted for use with survival data by Park et al. (2002) and Nguyen and Rocke (2002). This approach determines linear combinations of the features which are highly correlated with survival and orthogonal to each other. In Bastien's method, the weights in the first PLS component equal the regression coefficients of the univariate PH models. This is similar to the compound covariate method of Radmacher et al. (2002).

Proportional hazards models can also be built up in a stepwise iterative manner using boosting to select features for inclusion. The boosting idea is to iteratively generate models based on minimizing a sum of weighted residuals. The weights are adaptively modified to overweight cases which have large residuals. There are several approaches to PH model boosting (Hofner et al. 2014; Binder and Schumacher 2008), but in the example to follow, we have used the CoxBoost method.

Several authors have developed methods to identify optimal sets of genes which together provide good discrimination of the classes. These algorithms are generally very computationally intensive, and there is little evidence that they provide an improved prediction, though they have not been evaluated on enough real datasets (Lai et al. 2006; Staiger et al. 2013). Some of these methods are so computationally intensive that they have been applied using feature selection on the full dataset and then cross-validation for evaluation using those features. This partial cross-validation approach yields estimates of accuracy that is highly biased.

16.2.2 Aggregating Survival Trees

Segal (1998) introduced the concept of survival trees for regression and Hothorn et al. (2004, 2006) and Radespiel-Tröger et al. (2003) studied aggregation of decision trees to predict survival. A tree is built by successively splitting the cases into two groups based on the feature which is most associated with survival for the patients at that node. Nodes which are no longer split because of the stopping criterion called leafs of the tree. The stopping criterion may be based on the number of cases in a node, the depth of the path leading to the node, or the lack of effectiveness of splitting that node using the best single feature. Once the tree is constructed, prediction is performed for a new case by identifying the leaf that the case would be classified in and computing the Kaplan–Meier estimate for the training cases associated with that leaf.

Bagging of survival trees is carried out by taking B bootstrap samples of cases and computing a survival tree for each of the samples. Prediction of the probability of survival beyond time t for a new case is determined for each tree and averaged. Random forests are a variant of bagging survival trees in which only a sample of covariates is considered for splitting at each node of each tree. The number of covariates used is a pre-specified parameter. The R package party (Hothorn et al. 2006) can be used for building survival forests. One limitation of survival forests is that they are difficult to interpret because they may involve 1000 trees. They also are data hungry and contain many tuning parameters.

16.2.3 Neural Networks

Neural networks (NNs) are nonlinear regression methods that have infrequently been used for prediction survival risk but have been found to be of value in other fields. The simplest real NNs have three layers. The input layer has one node for each candidate prediction feature, and each node is connected to all the nodes of the intermediate "hidden" layer. The output layer contains a single node, and it is connected to all of the nodes of the hidden layer. The value of a node i in the hidden layer is often taken to be $H_i = 1/(1+\exp(-\boldsymbol{\alpha}_i'\mathbf{X}))$ where $\mathbf{X}$ denotes the vector of candidate predictors (values

of the input nodes) and $\boldsymbol{\alpha}$ denotes the vector of weights specific to hidden node i, one weight for each of the input features. Hence, each hidden node computes a different linear combination of inputs and transforms the value of the linear combination in a common nonlinear manner using the nonlinear "transfer function." The value of the output node is determined by the value of a nonlinear "activation function" $1/(1+\exp(-\boldsymbol{\beta}'\mathbf{H}))$ where $\mathbf{H}$ denotes the vector of values of the nodes in the hidden layer.

The value of the output node is taken as the predicted probability of survival beyond a pre-specified landmark time t. In training the network, cases with survival censored before t are omitted. Neural networks perform best with a very large training set. The number of parameters corresponding to the linear combination weights is p * m+m where p denotes the number of inputs and m is the number of hidden nodes. The number of hidden nodes is often kept small, but even then, fitting the network must generally be the second step of a feature selection process in order to keep p tractable. Otherwise, the data will be grossly overfit. One can aggregate NNs, sampling the candidate features to use for individual NNs in the manner of random forests. A large number of NNs can be built, and the predicted survival probability is a weighted average of the predictions of individual NNs, weighted by the accuracy of individual NNs. An alternative approach for limiting the number of input nodes is to use principle components as candidate predictors. Sargent (2001) compared prediction accuracy of ANN models to logistic regression models on 28 datasets with relatively small numbers of candidate predictors. ANN models were superior in some cases, but generally for the largest sample sizes, the two approaches gave similar results.

16.2.4 Clustering for Survival Risk Prediction

Unsupervised clustering can be used in several ways for survival risk prediction. Clinical investigators often use hierarchical clustering of the cases in order to identify clusters with relatively homogeneous gene expression signatures. They then compute Kaplan–Meier estimates of the survival distribution for each cluster as a way of establishing the clinical relevance of the clustering. This approach can be adapted for use as a survival risk prediction method if a supervised classifier is trained to assign new cases to the clusters. There is nothing special about using hierarchical clustering rather than some other method.

16.3 Accuracy Indices

Graf et al. (1999), Schumacher et al. (2003, 2007), and Bovelstad and Borgan (2011) have reviewed some of the measures of predictive accuracy for survival risk models. One index is the normalized predictive log partial likelihood which is shown in expression (16.2). The partial likelihood for case i is the model-based probability that case i fails at time t given that someone in the risk set at time t fails. This varies between 0 for a useless model to 1 for a perfect model. Because of the use of logarithms, however, this measure is not very intuitively interpretable and single cases that are not well predicted dominate the average.

Several measures of explained variation for survival data have been described (Korn and Simon 1990) One R^2 measure for PH models is of the form

$$R^2 = 1 - \exp\left(-\frac{2}{D}\left[\log(L(\hat{\beta})/L(0))\right]\right) \tag{16.3}$$

where L denotes the partial likelihood. This doesn't have all of the properties of the R^2 measure for normal linear models but is does provide an index in the interval [0, 1].

An alternative measure of the discriminatory power of a model is the c concordance index generalized for censored survival outcomes (Gonen and Heller 2005). For a PH model, the concordance index is empirical probability that the predictive index for case i is greater than that of case j given that the survival of case j is longer than that of case i. In computing this empirical probability, pairs in which both survivals are censored or in which the survival of one case is censored earlier than the uncensored survival of the other case are excluded.

For binary classification models (e.g., diseased or not diseased), the receiver operating characteristic (ROC) curve is a commonly used measure of prediction accuracy. If we have a classification model which provides a quantitative output, we can define sensitivity and specificity of the model for any specified value of a cut-point of the model output. The cut-point separates the predictions we consider as "disease" versus those we take as normal. The ROC curve is a graph of sensitivity on the vertical axis versus 1-specificity on the horizontal axis. The ROC concept has been generalized for survival data (Heagerty et al. 2000).

Sensitivity and specificity can be defined for predicting the probability of survival beyond a specified time t. Sensitivity is the probability of predicting survival beyond t for cases whose true survival times are greater than t. Sensitivity can be estimated in the following way:

$$\begin{aligned}\text{Sensitivity(c;t)} &= \Pr[\hat{S}(t) \geq c | T \geq t] \\ &= \frac{\Pr[T \geq t | \hat{S}(t) \geq c] \Pr[\hat{S}(t) \geq c]}{\Pr[T \geq t]}\end{aligned}$$

where T is the true unknown survival time and c is a cut-point. The first factor in the numerator can be estimated from the Kaplan–Meier curve for patients with a model predicted survival no less than c. The second factor in the numerator is estimated by the proportion of cases which have a model predicted survival no less than c. The denominator can be estimated from the Kaplan–Meier curve for all patients. The specificity at cut-point c and time t can be computed similarly. The ROC time-dependent curve is constructed by plotting sensitivity versus one minus specificity as a function of cut-point c for a fixed t. The area under the ROC curve can be computed for all t, and these AUC values plotted as a function of t.

The time-dependent ROC curve can also be constructed using pre-validated (Hofling and Tibshirani 2008) predictive values of $\hat{S}(t)$ as described below for calibration curves in the section on "Removing Over-fitting Bias".

The predictive accuracy of survival risk models can also be measured by the Brier score (Graff et al. 1999; Gerds and Schumacher 2006). The Brier score at time t is

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\hat{S}(t|X_i)^2 I(t_i \le t \wedge \delta_i = 1)}{\hat{G}(t_i)} + \frac{\left(1-\hat{S}(t|X_i)\right)^2 I(t_i > t)}{\hat{G}(t)}\right] \quad (16.4)$$

where t_i is the survival time of patient i with censoring indicator δ_I. I is the indicator function. $\hat{G}$ is the Kaplan–Meier estimate of the censoring distribution used to remove a large sample censoring bias. In formula (16.4), n is the number of evaluable cases, that is, cases which either survived beyond t or died earlier than t. The Brier score can be thought of as an estimate of the expected squared difference between the probability being estimated (surviving past t) and the outcome in the data (binary indicator of whether the patient survived past t). n denotes the sample size excluding the patients who were censored before time t.

16.3.1 Removing Over-Fitting Bias

The estimate of accuracy index computed on the same training set that was used to fit the model is called the re-substitution estimate. It is known that the re-substitution estimates of accuracy are optimistically biased, sometimes dramatically so. If there is a separate validation set, then all steps of model development, including feature selection, should be performed only using training set data. The model developed on the training set can then be used to predict survival outcomes for the patients in the validation set, and the accuracy indices are computed using only validation set cases. Dobbin and Simon (2011) discuss optimally splitting a dataset into training and validation portions when a fully independent validation set is not available.

An alternative to splitting a dataset into training and validation portions is to use a re-sampling procedure to estimate the bias of the re-substitution estimate of the accuracy index. This estimate is then used to adjust and de-bias the re-substitution estimate. For example, using the nonparametric bootstrap, B bootstrap sample is selected and withheld as a validation set. The model is developed on the non-withheld samples. The re-substitution estimate of the accuracy index is computed, and an unbiased estimate is computed based only on the accuracy of the model in predicting the withhold samples. The difference of these two estimates is an estimate of the bias. This bias estimate is averaged over the B bootstrap samples.

In cases where the withhold sample of cases is too small to estimate the statistic of interest, the pre-validation approach can be used. A K-fold cross-validation is performed, and a prediction is made for each withheld case. When the folds are completed, the predicted values, one per case, are combined and the statistic of interest, e.g., area under a time-dependent ROC curve, is computed using all of the predicted values. A prediction is made once for each case, and it is made using a model with that case omitted. But K models, containing possibly different features, are used for the predictions. This method was studied by Hofling and Tibshirant (2008) and was used by Simon et al. (2011) for evaluating the predictive accuracy of high-dimensional survival risk models and in the cross-validated adaptive signature design of Freidlin et al. (2010) for finding predictive treatment selection signatures.

16.3.2 Calibration of Survival Models

Evaluation of model accuracy is often composed of two components: the discriminatory power of the model and the calibration of the model. A model is considered well calibrated if it yields proper forecasts. For example, for the subset of men predicted to have a 30% chance of surviving 5 years, about 30% of them actually survive 5 years. For a given time t, let $\hat{S}(t; X)$ denote the predicted probability of surviving t years for a patient with covariate vector X. For any value
s ε (0, 1), let

$$\Omega(s) = \{X : |S(t, X) - s| < \varepsilon\}.$$

That is, Ω(s) denotes the set of covariate vectors X which have model predicted survival within epsilon of the selected value s. Let KM(t; s) denote the Kaplan–Meier estimate of the probability of surviving beyond time t, for the set of patients whose X vectors are included in Ω(s). The survival risk model is well calibrated if KM(t; s) ≈ s for all s in (0, 1). The calibration curve is a plot of s on the horizontal axis versus KM(t; s) on the vertical axis. With good calibration, the curve follows the 45° line.

It is best to compute the calibration curve using a separate test set. For each case in the test set, the value $\hat{S}(t; X)$ is computed for a pre-specified t. Those values then are used for computing the calibration curve using only cases in the validation set.

A pre-validated calibration curve can be constructed in the following way for datasets which are too small to split into training and validation. Using cross-validation, omit one or more cases. Fit the model for the remaining cases and compute $\hat{S}(t; X)$ for the cases having been omitted. Repeat this for all folds of the cross-validation. At that time, we have a predicted $\hat{S}(t; X)$ for all cases, and prediction for each case was made using a model developed on a training set that the case was omitted from. These $\{\hat{S}(t; X)\}$ values are called pre-validated (Tibshirani). The calibration curve is then computed using all the cases but with the pre-validated $\hat{S}(t; X)$ values.

16.3.3 Example

To illustrate some of these methods, we have used them to analyze data on survival of diffuse B cell lymphoma reported by Rosenwald et al. (2002). There were 414 patients and 165 deaths. Gene expression was evaluated on an Affymetrix microarray with 24,841 probesets. Overall survival was used as the endpoint. This analysis of one dataset is not meant to provide an adequate comparison of the methods. For a more extensive comparison, see Schumacher et al. (2007) and van Wieringen et al. (2009).

We evaluated several survival risk methods based on the PH model. We applied L1 penalized PH regression directly to the data with all 24,841 features although the number of features used in modeling could have been reduced using the methods of Tibshirani et al. (2012). We also performed L2 penalized PH regression. In order to obtain a model without too many variables, however, we included the first 100 principle components in the L2 penalized model instead of the 24,841 features. We also developed a partial least squares PH model using the first PLS component. We found that the package coxpls crashed the RStudio sessions we were running so we developed the PLS model on the first 100 principle components for dimensionality reduction. We computed the first PLS component and fitted a Cox PH model using that component using the survival package. The first PLS component has weights corresponding to the regression coefficients of single principle component PH models.

We also developed a supervised principle component PH model using the superPC package. We used only a single supervised principle component. It is the first principle component of the probesets which themselves are "significant" (nominal z statistic greater than 1.96 in absolute value) in univariate PH models. Finally, we developed a boosted PH model by applying the CoxBoost package directly to the probeset data.

Table 16.1 Survival models developed for diffuse large B cell lymphoma data

Model	Pre-validated concordance	Predictive partial likelihood	Pre-validated brier score at 3 years
L1 penalized PH	0.559	−884.5	0.314
L2 penalized PH on PCs	0.697	−912.4	0.213
Partial least squares on PCs	0.709	−912.3	0.200
Supervised PC	0.626	−870.2	0.290
CoxBoost PH	0.682	−912.7	0.223

To evaluate these models, we performed fivefold cross-validation. We computed the pre-validated concordance statistic, the predictive log partial likelihood, and the pre-validated Brier score for the probability of surviving beyond 3 years. Results are shown in Table 16.1. With regard to the concordance measure, the L2 penalized model based on pc's and the PLS model based on the pc's perform best.

With regard to predictive likelihood, the L2 penalized model, the PLS model and the CoxBoost model do best. The null log-likelihood is −956.11. Hence, the R^2 measure of expression (16.2) for the L2 penalized model is 0.41.

Finally, with regard to the Brier score at 3 years, the L2 penalized model, the PLS model, and the CoxBoost models again perform best. The L1 penalized model based on the original 24,841 probesets and the supervised pc model based on those probesets did not perform as well as the other models.

Figure 16.1 is pre-validated calibration plots for predicting the probability of 3-year survival for all five models. A properly calibrated probabilistic predictor should fall along the 45° line. The horizontal axis is the predictive probability of survival for beyond 3 years. The vertical axis is the proportion of patients who survive beyond 3 years. The L1 penalized PH model and the supervised PC model are seen to be poorly calibrated. The other three models are well calibrated, particularly the CoxBoost PH model.

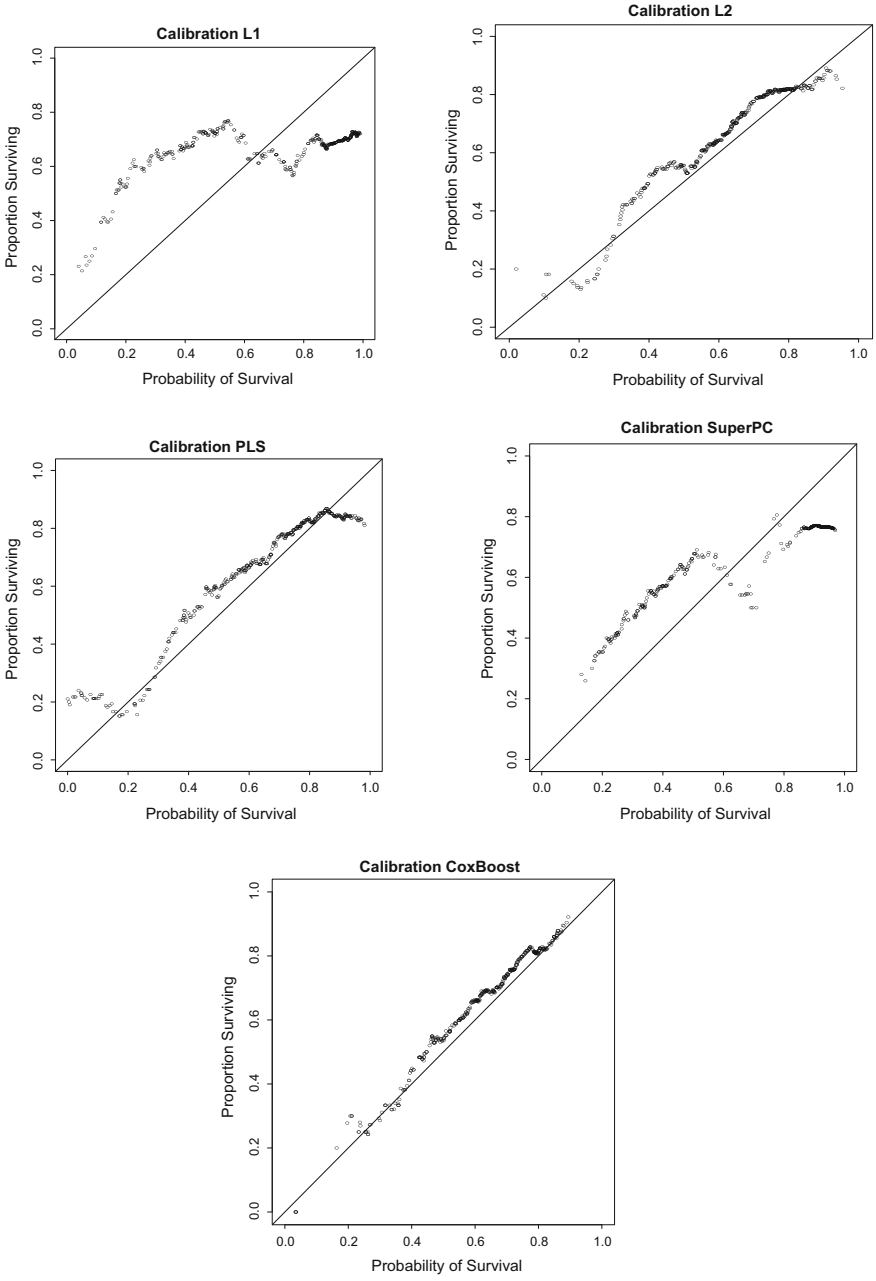


Fig. 16.1 Calibration of survival models developed for diffuse large B cell lymphoma data

References

Bair, E., & Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, *2*, 511–522.

Binder, H., Schumacher M (2008) Allowing for mandatory covariates in boosting estimation of sparse high dimensional survival models. *BMC Bioinformatics, 9*, 14.

Bovelstad, H. M., & Borgan, O. (2011). Assessment of evaluation criteria for survival prediction from genomic data. *Biometrical Journal, 53,* 202–216.

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics, 4,* 31.

Freidlin, B., Jiang, W., & Simon, R. (2010). The cross-validated adaptive signature design. *Clinical Cancer Research, 16*(2), 691–698.

Gerds, T. A., & Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal, 48,* 1029–1040.

Gönen, M., & Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika, 92,* 965–970.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine, 18*(17–18), 2529–2545.

Hastie, T., & Tibshirani, R. (2004). Efficient quadratic regularization for expression arrays. *Biostatistics*, *5*, 329–340.

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics, 56,* 337–344.

Hothorn, T., Benner, A., Lausen, B., & Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in Medicine*, *23*, 77–91.

Höfling, H., & Tibshirani, R. (2008). A study of pre-validation. *The Annals of Applied Statistics, 2,* 643–664.

Hofner, B., Mayr, A., Robinzonov, N., & Schmid, M. (2014). Model-based boosting in R: A hands-on tutorial using the R package mboost. *Computational Statistics, 29,* 3–35.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van der Laan, M. (2006). Survival ensembles. *Biostatistics*, *7*, 355–373.

Korn, E. L., & Simon, R. (1990). Measures of explained variation for survival data. *Statistics in Medicine, 9,* 487–503.

Lai, C., Reinders, M. J., van't Veer, L. J., & Wessels, L. F. (2006). A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics*, *7*, 235.

Nguyen, D. V., & Rocke, D. M. (2002). Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics, 18,* 1625–1632.

Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., et al. (2006). Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor–positive breast cancer. *Journal of Clinical Oncology, 24,* 3726–3734.

Park, P. J., Tian, L., & Kohane, I. S. (2002). Linking expression data with patient survival times using partial least squares. *Bioinformatics*, *18*, S120–S127.

Radespiel-Tröger, M., Rabenstein, T., Schneider, H. T., & Lausen, B. (2003). Comparison of tree-based methods for prognostic stratification of survival data. *Artifical Intelligence in Medicine*, *28*, 323–341.

Radmacher, M. D., McShane, L. M., & Simon, R. (2002). A paradigm for class prediction using gene expression profiles. *Journal of Computational Biology, 9,* 505–511.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine, 346,* 1937–1947.

Sargent, D. J. (2001). Comparison of artificial neural networks with other statistical approaches. *Cancer, 91*(S8), 1636–1642.

Segal, M. R. (1998). Regression trees for censored data. *Biometrics*, *48*, 35–47.

Schumacher, M., Graf, E., & Gerds, T. (2003). How to assess prognostic models for survival data: A case study in oncology. *Methods Archive, 42,* 564–571.

Schumacher, M., Binder, H., & Gerds, T. (2007). Assessment of survival prediction models based on microarray data. *Bioinformatics, 23,* 1768–1774.

Simon, R. M., Subramanian, J., Li, M. C., & Menezes, S. (2011). Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Briefings in Bioinformatics, 12,* 203–214.

Staiger, C., Cadot, S., Györffy, B., Wessels, L. F., & Klau, G. W. (2013). Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Frontiers in Genetics*, *4*.

Subramanian, J., & Simon, R. (2010). Gene expression–based prognostic signatures in lung cancer: Ready for clinical use? *Journal of the National Cancer Institute, 102,* 464–474.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine, 16,* 385–395.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., et al. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74,* 245–266.

Van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., Van't Veer, L. J., & Wessels, L. F. A. (2006). Cross-validated cox regression on microarray gene expression data. *Statistics in Medicine*, *25*, 3201–3216.

van Wieringen, W. N., Kun, D., Hampel, R., & Boulesteix, A. L. (2009). Survival prediction using gene expression data: A review and comparison. *Computational Statistics & Data Analysis, 53,* 1590–1603.

Chapter 17
Validation, Multivariate Modeling, and the Construction of Heat-Map Prediction Matrices for Survival in the Context of Missing Data

Shankar S. Srinivasan, Albert Elion-Mboussa and Li Hua Yue

17.1 Introduction

Non-interventional trials or registries allow some latitude in the reporting of observations and procedures by site investigators, leading to a larger degree of missing data than controlled clinical trials. However, controlled clinical trials (CCT), which tend to have more complete data, have an extensive set of inclusion and exclusion criteria which make their patient populations highly selective and unrepresentative of the typical patient presenting at a clinic. Furthermore, sites at CCTs tend to be research or academic sites rather than regular community sites. However, registries typically do not work under these constraints and provide ideal data for the development of prognostic models for the typical patient, whereas the generalizability of models derived from CCTs maybe in question (Unger et al. 2014; Vist et al. 2008; Townsley et al. 2005). Hence the objective of this chapter is to perform prognostic analyses within the context of real-world non-interventional trials while circumventing the drawback of data incompleteness.

The registry through which we will illustrate the construction of prediction models, despite incomplete data, is the Connect® MM Registry (NCT01081028) (http://clinicaltrials.gov/ct2/show/NCT01081028). This registry enrolled two cohorts. The first cohort has an adequate follow-up (median 33.5 months, N = 1493) for analysis, while analysis for the second cohort is pre-mature due to inadequate follow-up.

S. S. Srinivasan (✉) · A. Elion-Mboussa · L. H. Yue
Department of Biostatistics, Celgene Corporation, Summit, NJ, USA
e-mail: shsrinivasan@celgene.com

A. Elion-Mboussa
e-mail: aelionmboussa@lexpharma.com

L. H. Yue
e-mail: lyue@celgene.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_17

The Connect® MM Registry was designed as a prospective, observational, longitudinal, multicenter study of patients with newly diagnosed multiple myeloma (MM). There is no planned investigational agent, prescribed treatment regimen, or mandated intervention in this study. The treating physician determines the enrolled patient's therapy for newly diagnosed MM per his or her clinical judgment. Inclusion criteria were limited to patients who were newly diagnosed with symptomatic MM within 2 months of enrollment, age ≥ 18 years, willingness, and ability to sign informed consent and an agreement by the patient to complete patient questionnaires alone or with minimal assistance. There were no exclusion criteria. We retrospectively applied the typical collection of additional CCT inclusion and exclusion criteria (such as M-protein ≤ 1.0 g/dL, creatinine > 2.5 mg/dL, ANC $\leq 1.5 \times 10^9$/L, hemoglobin ≤ 8 g/dL, AST/ALT > 3 UNL, platelets $\leq 75 \times 10^9$/L.) to the data from this registry and found that 40% of the patients in the registry would not have been eligible for a CCT, reflecting that the registry had adequate non-selectivity making it prime data for prognostic analysis (Shah et al. 2017). Further supporting, the utility of the data was the dominance of community sites (81.1%) while the prevalence of academic and government investigational sites was not insignificant (17.6% and 1.3%, respectively) (Rifkin et al. 2015). We evaluated the registry baseline data against the National Comprehensive Cancer Network's suggested diagnostic workup for multiple myeloma and found that allowing physician discretion in diagnostic data to be collected, as is usually done for non-interventional registries, had led to a small to moderate, and sometimes a large extent of incomplete data (Rifkin et al. 2015). So, while this large registry database represents an ideal cohort for prediction models, the incompleteness of the data makes the development of such models challenging and interesting.

We intend to explain the methodology behind the development of an appealing heat-map representation predicting early and late survival along with the associated model building and validation. Such a representation will help facilitate patient—physician interactions at the diagnosis of multiple myeloma. Similar heat-map representations, developed using CCT data, have been found useful in the context of rheumatoid arthritis and ankylosing spondylitis (Vastesaeger et al. 2011; Vastesaeger et al. 2009). This striking representation is both a color-coded visual summary as well as a numeric summary of predictions about survival as related to an identified set of prognostic factors. The heat-map representation based on the Connect® MM Registry with complete clinical background and interpretation has been published (Terebelo et al. 2017) for early survival. A manuscript is under development for late survival using additional follow-up. The beta version of that late survival model is presented in this manuscript while the methods described reflect some enhancements in the updated analysis.

Model building is illustrated for a discrete and for a survival endpoint in the context of missing data. The discrete endpoint being considered is the mortality within a 180-day period (6 months). The survival endpoint being considered is overall survival, from which the 1-, 2-, 3-, and 4-year survivals are of interest and are derived. The decision for separate analyses for early and later mortality followed from the perception that early mortality may have a different etiology than later mortality with

co-morbidities dominating the former and disease factors more relevant for the latter. Section 17.2 will elaborate, for both endpoints, on univariate screening from a totality of all identified relevant baseline characteristics, multiple imputation of the missing data, and variable selection in the context of imputed data. Section 17.3 will provide, for the discrete endpoint, the combined inferences, the heat-map representation, and the internal and external validation details. Section 17.4 will provide similar details for the later mortality. The chapter will conclude with a discussion section.

17.2 Multivariate Modeling

Multivariate modeling starts with univariate screening to reduce the number of predictors and follows up with variable selection before developing the predictive model. This section covers these steps in the context of missing data. SAS code is provided in text for the discrete endpoint. The code is similar for the survival endpoint and uses SAS PROC PHREG code instead of PROC LOGISTIC code. This survival code is not provided.

17.2.1 Univariate Screening

Univariate analyses were conducted with the intent of determining the degree of missingness on each predictor and statistical significance of the predictor in predicting the dependent measure. Variables significant at the 0.15 level and with less than 60% missing were screened in. Logistic regression analyses were conducted for the discrete variable of mortality within 180 days and Cox regression analyses were conducted on survival data. The code for the logistic regression follows, where d180 is the discrete dependent variable.

```
proc logistic data = Edeath descending;
    model d180 = &var/risklimits;
    ods output ParameterEstimates=&univ_est NObs = &univ_miss;
run;
```

Twenty-six variables were screened. Thirteen variables were screened through the logistic regression analyses, and seventeen variables were screened through Cox regression analyses. The average amount of missing data for the logistic regression screened variables was 9.23% and 13.1% for the Cox regression.

17.2.2 Multiple Imputation

Concern about the degree of missing data is mitigated by the number of imputed datasets we created. The relative efficiency *RE* of multiple imputation is given by

$$RE = (1 + \lambda/m)^{-1}$$

where λ is the fraction of missing information about the parameter being estimated and m is the number of imputed datasets (Little and Rubin 2002; Yuan 2000). The fraction of missing data will be roughly proportional to the average amount of missing data reported earlier. For three imputations, the *RE* is 0.9375 and 0.8571 for missing fractions of 20% and 50%, respectively. For the intended ten imputations, the *RE* increases to 0.9804 and 0.9524, respectively.

Rubin's imputation framework (Little and Rubin 2002) was used for the analysis. This involves assuming an imputation model, then obtaining the predictive distribution of the missing data conditional on observed data and distribution parameters, and then producing multiple imputed datasets using the predictive distribution (Patricia and Steven 2014). Analysis under multiple imputation is robust under less restrictive assumptions of missing at random (MAR) compared to the case-wise deletion of data records with any data missing on any predictor. Further case-wise deletion of data missing on any variable leads to considerable loss of information on other collected variables. The imputation model was the fully conditional specification (FCS) as recommended in (van Buuren 2007). All variables (including those screened out) were used in the imputation model to extract all information on the missingness of the predictors contained in the dataset. Ten imputations were generated. The SAS code for the FCS method follows.

```
proc mi data=os nimpute=10 seed=5122017 out=osm;
    class agen hispan bmi issstagen ecogn …;
    fcs logistic(agen hispan bmi issstagen ecogn …);
    var agen hispan bmi issstagen ecogn …;
run;
```

17.2.3 Variable Selection

To find the candidate multivariate models, the imputed datasets were stacked on top of each other and the multivariate logistic and Cox regressions were run using underweighted observations, with the underweighting proportional to the number of imputed datasets and to the degree of missingness, an approach recommended as one of the reasonable approaches in (Wood et al. 2008). The predictors used were those screened in, in Sect. 17.2.1. The SAS code for the logistic model requesting all possible models follows. The weight = (1 − f)/(# of imputations) where f is the average fraction of missing data.

```
proc logistic data = Edeathm2 ;
    model d180 (event = 'yes') = agen issstagen mhecogynn imwg_risk
    mhdiabn mhhyn calcium creat plat_ct caref mobf gp_17p_ad novelf/
    selection = score details lackfit ;
       weight wt;
run;
```

Selection = score in SAS provides the score statistic for all possible models. The difference in score statistics between models is a chi-squared distribution with degrees of freedom given by the difference in the number of variables in the models. Starting with best 1 variable model, we moved in 1 variable increment to the best k variable model, till the incremental score statistic is less than the critical value obtained as the 0.1-level Wald X^2 chi-square value for 1 degree of freedom. Several models with score statistics in the neighborhood of that for the best k variable model were then considered as candidate models. For each candidate model, multivariate logistic/Cox regression was fit on each of the ten imputed datasets, and the average Bayesian information criterion (BIC) value was calculated. The final multivariate model was selected as the candidate model with the minimum average BIC among models judged to be clinically appropriate. Seven variables were selected through the stacked weighted logistic regression, and eleven variables were selected through the variable selection process using Cox regression. The next section starts by using Rubin's method (Little and Rubin 2002) to combine the inferences for the seven variable logistic regressions applied to each imputed dataset.

17.3 The Discrete Case

Model building continues in this section for the discrete endpoint with the computation of the logistic model using imputed datasets, the development of an intuitive representation of predictions from the model and internal and external validation.

17.3.1 Combining Inferences

By Rubin's result (Little and Rubin 2002), the estimate of a parameter of interest is the average of estimates from each imputed dataset. Such an estimate is efficient and unbiased under MAR assumptions. The separate estimates and the combined inferences were obtained using the following SAS code for the seven selected variables

```
proc logistic data=Edeathm2;
model d180 (event = 'Yes') = agen issstagen mhecogynn mhhyn creat
plat_ct mobf /risklimits details lackfit covb;
by _Imputation_;
ods output ParameterEstimates=lgparms CovB=lgcovb;
run;

proc mianalyze parms=lgparms
covb=lgcovb;
modeleffects Intercept agen issstagen mhecogynn mhhyn creat plat_ct
mobf;
ods output ParameterEstimates=est1;
run;
```

Table 17.1 Logistic regression analysis of baseline characteristics predictive of mortality within 180 days

Characteristic	Odds ratio	95% Confidence interval	P-value
Age (>75 vs. ≥75)	1.70	(1.09, 2.67)	0.020
ECOG performance score (≥2 vs.<2)	3.89	(1.67, 9.05)	0.002
History of hypertension	1.96	(1.19, 3.22)	0.008
ISS disease stage (III vs. I and II)	1.85	(1.18, 2.90)	0.007
Renal insufficiency (Serum Creatinine > 2 mg/dL)	1.59	(0.98, 2.60)	0.062
Platelet count (<150 × 10^9/L vs.>150 × 10^9/L)	2.29	(1.49, 3.53)	<0.001
Mobility from EQ5D	2.42	(1.48, 3.94)	<0.001

The output dataset est1 above contains the estimates of the intercept parameter α and the regression coefficients β's for each predictor x_i in the logistic model given by

$$\pi(\mathbf{x}) = \frac{\exp\left(\alpha + \sum_{i=1}^{p} \beta_i x_i\right)}{1 + \exp\left(\alpha + \sum_{i=1}^{p} \beta_i x_i\right)}$$

where $\pi(\mathbf{x})$ is the probability of the event corresponding to a vector of predictor values $\mathbf{x}$ (Stokes et al. 2012). Exponentiation of the parameter estimates and confidence limits provides the odds ratios for a one-point increment in the predictor variable. All variables, except for mobility, were dummy coded as 0 and 1 as they were dichotomized variables. Mobility is ordinal and takes three levels from 0 to 2 and the odds ratio represents, on average, the change in odds for every increase in the level of mobility. Table 17.1 provides a summary of inferences from the final logistic model using multiple imputation. The odds ratio of 1.70 implies that the odds of mortality within 180 days for those patients with age > 75 are 1.7 times that for those ≤75 years. Similar interpretations apply to other characteristics in Table 17.1.

17.3.2 Heat-Map Prediction Matrices

The heat-map prediction matrices are designed to show less favorable outcomes in the bottom left corner and more favorable outcomes toward the top right corner of the matrix. Follow our notes on the steps to achieve this objective by examining Fig. 17.1. We start by ordering the variables on importance which is assessed by multiplying the odds ratio by (# of predictor levels—1). For instance, mobility was assessed most relevant to the matrix in Fig. 17.1 as $2.42 \times (3 - 1) = 4.84$ is the largest computed value. We start by placing this in the largest row header of the matrix. ECOG status is the next most important and gets to be the largest column header of the matrix. The third most relevant variable platelet count bifurcates the mobility header. The fourth most important variable of hypertension history bifurcates the ECOG header. Similarly, alternating between rows and columns populates row and column headers with all model predictors. The row header predictors have the predictor level with the favorable outcome on top. The column header predictors have the predictor level with the favorable outcome to the right. This generates a blank matrix with column and row headers. To populate these blank cells with the appropriate prediction, we generate every combination of the constituent levels of the predictors, map each

	ECOG ≥ 2			ECOG ≥ 2			ECOG < 2			ECOG < 2	
Mobility: No problem in walking about	12%	8%	PC > 150	7%	4%	ISS I, II	3%	2%	Creat ≤ 2	2%	1%
	18%	11%		10%	6%		5%	3%	Creat > 2	3%	2%
	20%	13%		12%	7%	ISS III	6%	4%	Creat ≤ 2	3%	2%
	29%	19%		17%	11%		10%	6%	Creat > 2	5%	3%
	24%	16%	PC ≤ 150	14%	9%	ISS I, II	8%	5%	Creat ≤ 2	4%	2%
	34%	23%		21%	13%		12%	7%	Creat > 2	6%	4%
	37%	26%		23%	15%	ISS III	13%	8%	Creat ≤ 2	7%	4%
	48%	36%		32%	22%		19%	12%	Creat > 2	11%	7%
	Hypertension = Yes			Hypertension = No			Hypertension = Yes			Hypertension = No	
Mobility: some problem in walking about	25%	16%	PC > 150	15%	9%	ISS I, II	8%	5%	Creat ≤ 2	4%	3%
	35%	24%		21%	14%		12%	7%	Creat > 2	7%	4%
	38%	27%		24%	16%	ISS III	14%	9%	Creat ≤ 2	8%	5%
	50%	37%		34%	23%		20%	13%	Creat > 2	11%	7%
	43%	31%	PC ≤ 150	28%	19%	ISS I, II	16%	10%	Creat ≤ 2	9%	6%
	55%	42%		38%	27%		24%	16%	Creat > 2	14%	9%
	59%	46%		42%	30%	ISS III	27%	18%	Creat ≤ 2	16%	10%
	69%	57%		54%	40%		37%	26%	Creat > 2	23%	15%
	Age > 75	Age ≤ 75		Age > 75	Age ≤ 75		Age > 75	Age ≤ 75		Age > 75	Age ≤ 75
Mobility: confined to bed	45%	32%	PC > 150	29%	19%	ISS I, II	17%	11%	Creat ≤ 2	10%	6%
	56%	43%		40%	28%		25%	16%	Creat > 2	14%	9%
	60%	47%		43%	31%	ISS III	28%	18%	Creat ≤ 2	16%	10%
	71%	58%		55%	42%		38%	27%	Creat > 2	24%	16%
	65%	52%	PC ≤ 150	49%	36%	ISS I, II	32%	22%	Creat ≤ 2	20%	12%
	75%	63%		60%	47%		43%	31%	Creat > 2	28%	19%
	77%	67%		64%	51%	ISS III	47%	34%	Creat ≤ 2	31%	21%
	85%	76%		74%	62%		59%	45%	Creat > 2	42%	30%

Fig. 17.1 Prediction Matrix for Mortality within 180 days for newly diagnosed multiple myeloma

combination to the cells of the matrix, and compute the probability of early mortality for each combination. The estimated probabilities can then be easily inserted into the prediction matrix. SAS code to generate data for insertion into the section of the matrix where mobility = 0 (No problem in walking about) is provided in Appendix 1.

In Fig. 17.1, we move to smaller blocks within the larger blocks with factors which have succeeding smaller effects. Numeric values in the matrix are the estimated probability of mortality within 180 days. The chart is traffic color-coded—green for favorable outcomes, from yellow to red for less favorable outcomes. A color version of this figure is available in the online version of this book. An excel app facilitating the easy selection of combinations of predictor levels and the identification of the estimated probability has been developed and will be posted at a Celgene Corporation Registry Web site soon. The location of this Web site as well as further details will be presented at www.resourcetepee.com.

17.3.3 Internal Validation

Internal validation involves the splitting of the dataset into test and training samples. The model obtained in the training sample is evaluated in the test sample. Better estimates of validation indices are obtained when they are obtained through analysis of repeated random splits into test and training samples—A process called bootstrap re-sampling. The validation index we used to measure the predictive ability of our prognostic model was the Harrell's C-Index (Harrell 2001). This index is interpretable as a concordance probability—the probability that a randomly selected pair of patients, one with a poorer survival outcome than the other, will be correctly differentially identified based on inputting the two patient's baseline prognostic characteristics in the fitted model. To compute the index, we imported each of the ten imputed datasets into the R software and ran the following R code for each dataset for 100 bootstrap sample pairs

```
library("rms")
## Imputation # 1
f <- lrm(d180 ~ age+iss+ecog+hyptension+platcount+mobility+creatine,
data = impt1log, x=TRUE, y=TRUE)
validate(f, B=100, dxy = TRUE)
```

This R script provides the Somer's D statistic Dxy. The concordance probability for each imputation can be computed as C-Index = 0.5 * |Dxy| + 0.5. Training datasets have better predictive ability due to the possibility of over-fitting the model to the data. The training optimism adjusted concordance probability adjusts for this bias. In the multiple imputation context, we compute the concordance probability as the average of the adjusted concordance probabilities from each imputation. For the logistic model, the concordance probability is identical to the area under the receiver operating characteristic (ROC) curve for the model and confidence intervals can therefore be computed using expressions developed in (Hanley and McNeil 1982)

for this area under the curve. The percent reduction in the concordance probability for the test samples compared to the training samples was 2.53% for the logistic model indicating the unlikelihood of an over-fitted model. The training optimism adjusted concordance probability of the fitted logistic model was estimated at 74.3% (95% CI: 68.7, 80.0). A concordance probability significantly greater than 50% is indicative of a good predictive model.

17.3.4 External Validation

External validation is a measure of how well the model derived from our registry works for an independent external dataset. The external data we chose was that for the FIRST multiple myeloma clinical study (N = 1623). A description of this clinical study and efficacy and safety during this trial is reported in (Benboubker and Dimopoulos 2014). This study was a phase III, randomized, open-label, 3-arm study to determine the efficacy and safety of lenalidomide (Revlimid®) plus low-dose dexamethasone when given until progressive disease or for 18 four-week cycles versus the combination of Melphalan, Prednisone, and Thalidomide given for 12 six-week cycles in patients with previously untreated multiple myeloma who are either 65 years of age or older or not candidates for stem cell transplantation.

All seven variables which were significant predictors in the logistic model developed from the registry data were collected in the FIRST study. These variables as well as mortality within 180 days were extracted from the FIRST database. Then the probability of early mortality was computed for the FIRST data using the registry-derived logistic model and compared against actual outcomes in the FIRST study. This was achieved through the R package rms through the following

```
library(rms)
phat <- 1/(1+exp(-(-4.543656+(0.883258*logist$mobf
+0.673436*logist$mhhyn+1.359005*logist$mhecogynn+0.617535*logist$isss
tage+0.533151*logist$agen+0.830696*logist$plat_ct+0.466740*logist$cre
at))))
val.prob(phat, logist$dthbf180, xlab="Predicted Probability of Death
Before 180 Days ", ylab=" Actual Probability of Death Before 180 Days
", lim=c(0,1.0), legendloc = c(0.75,0.15), m= 30, cex = 0.7)
```

This R script produces Fig. 17.2. The concordance probability is 71.83% (95% CI: 66.2, 77.4) compares favorably to 74.3% in the internal validation supporting the portability of the derived model. This is despite differences between FIRST, a CCT, and the registry. The triangles in the graphic represent sets of 30 patients. The data plots under the equiangular line, likely, because FIRST had a healthier cohort due to restrictive inclusion and exclusion criteria. This would have led to lower actual than predicted probabilities.

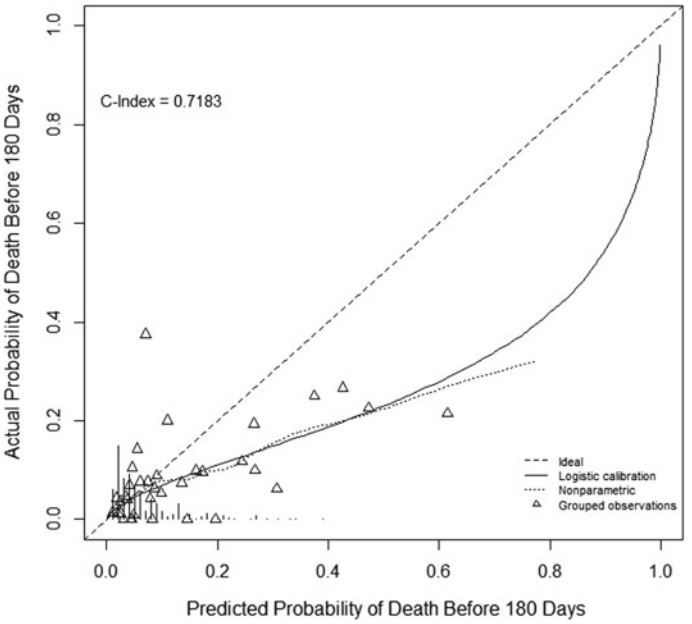


Fig. 17.2 External validation of the logistic model

17.4 The Survival Case

Model building continues in this section for the survival endpoint with the computation of the Cox model using imputed datasets, the development of an intuitive representation of predictions from the model and internal and external validation.

17.4.1 Combining Inferences

The separate estimates by imputation and the combined inferences were obtained using the following SAS code for the eleven selected variables

```
proc phreg data=Edeathm2;
model dur*death(0) = agen issstagen mhecogynn mhdiabn creat plat_ct
mhsolynn mobf gp_17p_ad gp_hyper_ad novelf
by _Imputation_;
ods output ParameterEstimates=lgparms CovB=lgcovb;
run;

proc mianalyze parms=lgparms
covb=lgcovb;
modeleffects  agen issstagen mhecogynn mhdiabn creat plat_ct mhsolynn
mobf gp_17p_ad gp_hyper_ad novelf;
ods output ParameterEstimates=est1;
run;
```

The output dataset est1 above contains the estimates of the regression coefficients β's for each predictor x_i in the Cox model given by

$$h(t, \mathbf{x}) = h_0(t) \exp\left(\sum_i^p \beta_i x_i\right)$$

where $h(t, \boldsymbol{x})$ is the hazard function at time t defined at a vector of predictor values $\mathbf{x}$ and $h_0(t)$ is the baseline hazard function. Exponentiation of the parameter estimates and confidence limits provide the hazard ratios and confidence limits for a one-point

Table 17.2 Cox regression analysis of baseline characteristics predictive of overall survival

Characteristic	Hazard ratio	95% Confidence interval	P-value
Age (>75 vs. ≥75)	2.01	(1.66, 2.44)	<0.001
ECOG performance score (≥2 vs. <2)	1.73	(1.09, 2.74)	0.019
History of diabetes	1.37	(1.11, 1.69)	0.003
Del (17P) from FISH and cytogenetic forms	1.55	(1.18, 2.04)	0.002
Hyperdiploidy	1.65	(1.22, 2.25)	0.002
Extramedullary plasmacytoma	1.47	(1.13, 1.91)	0.004
ISS disease stage (III vs. II vs. I)	1.31	(1.15, 1.50)	<0.001
Renal insufficiency (Serum Creatinine > 2 mg/dL)	1.46	(1.16, 1.83)	0.001
Platelet count (<150 × 10^9/L vs. >150 × 10^9/L)	1.47	(1.21, 1.79)	<0.001
Mobility from EQ5D	1.49	(1.24, 1.79)	<0.001
Novel therapy use ≥2 versus (0, 1)	0.78	(0.62, 0.98)	0.033

increment in the predictor variable. All variables in the Cox model, except for mobility and ISS stage, were dummy coded as 0 and 1 as they were dichotomized variables. Mobility and ISS are ordinal and take three levels and the hazard ratio represents, on average, the change in hazard for every increase in level. Table 17.2 provides a summary of inferences from the final Cox model using multiple imputation. The hazard ratio of 2.01 for age implies that the hazard of mortality for those patients with age >75 is 2.01 times that for those ≤ 75 years. Similar interpretations apply for other characteristics in the table.

17.4.2 Heat-Map Prediction Matrices

Heat-map prediction matrices featuring predictions of estimated survival beyond three years will be constructed using methods like those described in Sect. 17.3.2 for the discrete endpoint. Due to the large number of predictors in the Cox model, we split the prediction matrix into two panels by age group, a clinically important variable in multiple myeloma. These panels are in Figs. 17.3b and 17.4b. As with the discrete case, we order the variables by importance and then assign them by alternating the predictors between the rows and column headers. As before, the row header predictors have the predictor level with the favorable outcome on top. The column header predictors have the predictor level with the favorable outcome to the right. This generates a blank matrix with column and row headers. To populate these blank cells with the appropriate predictions, we generate every combination of the constituent levels of the predictors, map each combination to the cells of the matrix, and compute the probability of survival beyond three years for each combination. The SAS code to implement this starts with SAS PROC PLAN code as in the discrete case in Appendix 1 and generates a dataset covals containing the combinations of the levels of the predictors along with the mapping to cells in the matrix. However, the SAS code for the survival case differs from that of the discrete case in the way predictions are generated.

To generate the predictions, we use the methods in Allison (2010). The code below uses the covals dataset in the baseline statement of the SAS PHREG procedure to generate survival probabilities at every event time in the registry along with confidence intervals. To obtain the survival beyond three years, we retain the data records corresponding to event time closest to and less than the three-year time-point (1095 days). The prediction of survival beyond three years for each predictor combination is estimated as the average of the corresponding 3-year survivals from each of the imputations.

```
proc phreg data=Edeathm2;
   model dur*death(0) = agen issstagen mhecogynn mhdiabn creat
   plat_ct mhsolynn mobf gp_17p_ad gp_hyper_ad novelf /ties=efron;
   baseline out=a covariates=covals survival=s lower=lcl
   upper=ucl/nomean;
   by _Imputation_;
run;
```

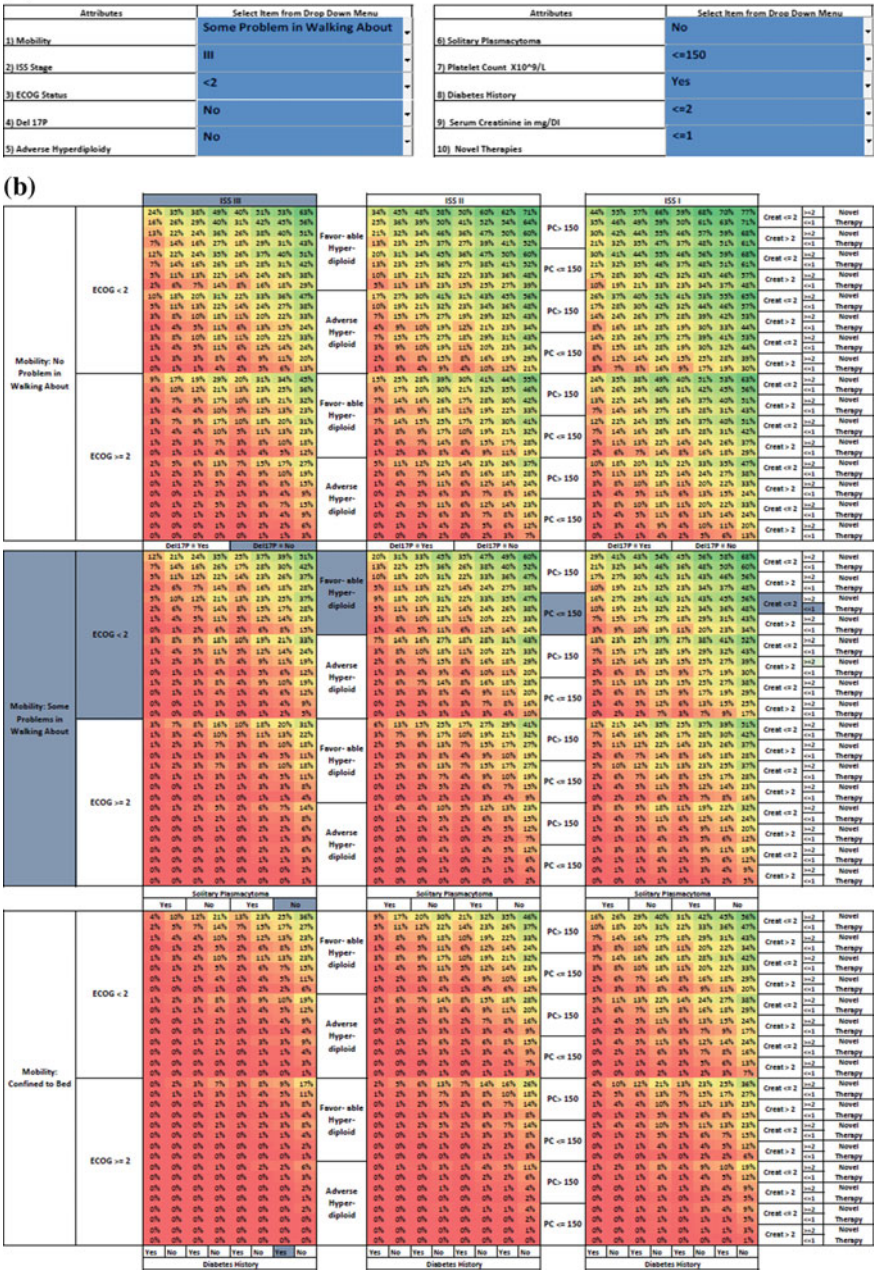


Fig. 17.3 **a** Excel app inputs to prediction matrix for elderly patients (age > 75) in (**b**). **b** Prediction Matrix of survival probability beyond three years for elderly patients

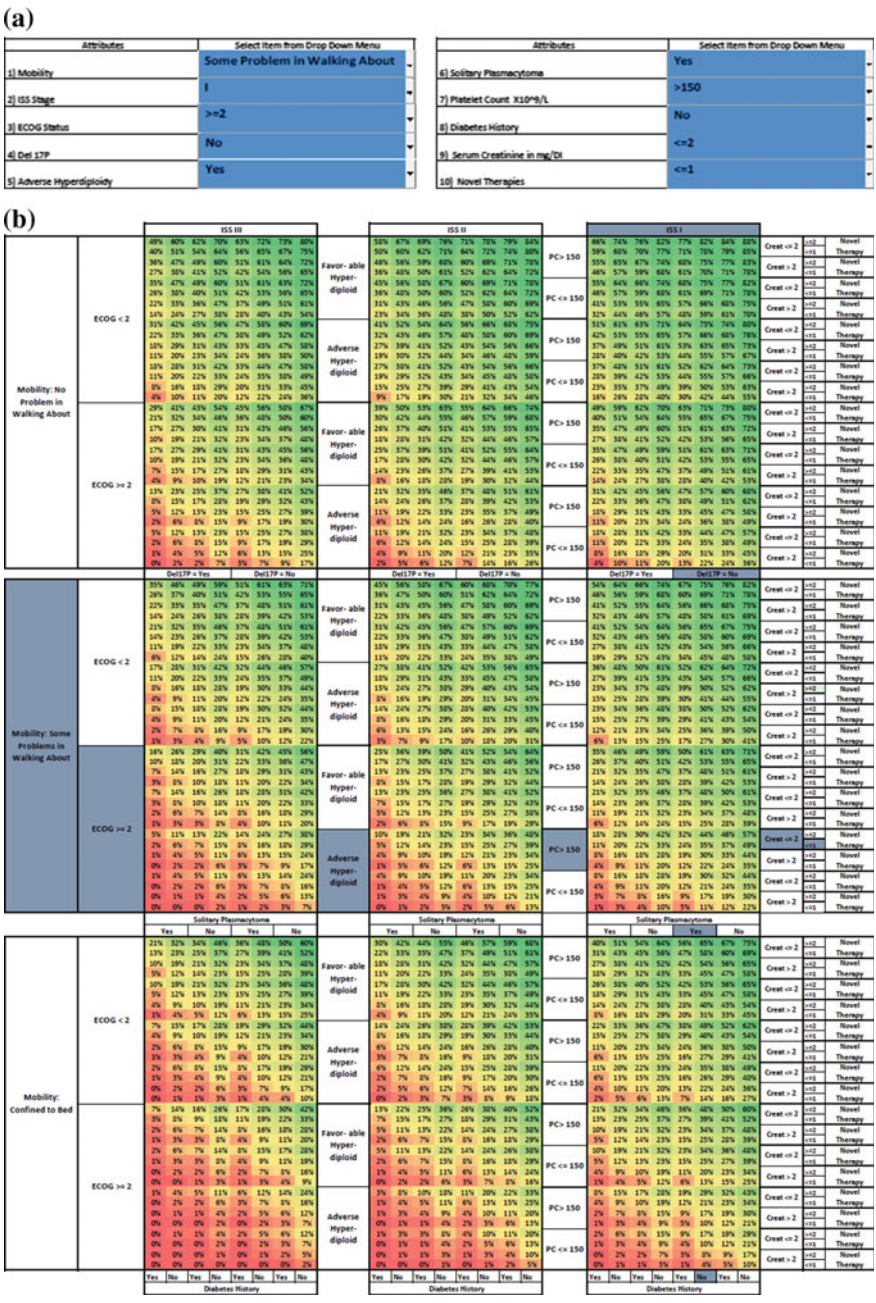


Fig. 17.4 **a** Excel app inputs to prediction matrix for younger patients in (**b**). **b** Prediction Matrix of survival probability beyond three years for younger patients (age ≤ 75)

In Figs. 17.3b and 17.4b, we move to smaller blocks within the larger blocks with factors which have succeeding smaller effects. Numeric values in the matrix are the estimated probability of survival beyond 3 years. The panels are traffic color-coded as in Fig. 17.1 for the discrete data. A color version of this figure is available in the online version of this book. An excel app facilitating the easy selection of combinations of predictor levels and the identification of the estimated probability has been developed. Instructions on how to access this app and apps for 1-year, 2-year, and 4-year survival will be posted at www.resourcetepee.com. Figure 17.3b is a screenshot of the panel for the elderly obtained from the inputs to the app in Fig. 17.3a.

For the inputs in Fig. 17.3a, the app highlights the appropriate row and column headers in Fig. 17.3b. Following the shaded columns and rows, we see that a patient with the characteristics in Fig. 17.3a has an estimated 17% probability of surviving beyond 3 years.

Figure 17.4b is a screenshot of the panel for the younger patients obtained from the inputs to the app in Fig. 17.4a above.

For the inputs in Fig. 17.4a, the excel app highlights the appropriate row and column headers in Fig. 17.4b. Following the shaded columns and rows, we see that a patient with the characteristics in Fig. 17.4a has an estimated 35% probability of surviving beyond 3 years.

17.4.3 Internal Validation

As for the discrete case, internal validation involved bootstrap re-sampling of a 100 test and training datasets and the computation of concordance probabilities. To compute this concordance index, we imported each of the ten imputed datasets into the R software and ran the following R code:

```
library("rms")
## Imputation # 1
f <- cph(formula=Surv(dur,death) ~
age+iss+ecog+diabetes+creatine+platcount+plasmacytoma+mobility+novel+
del17p+hyperploid, data = impt1dt, x=TRUE, y=TRUE, surv = T)
validate(f,B=100, dxy =TRUE)
```

Using computations like those for the discrete case, the percent reduction in the concordance probability for the test samples compared to the training samples was 0.94% for the Cox model indicating the unlikelihood of an over-fitted model. The training optimism adjusted concordance probability of the fitted Cox model was estimated at 69.5% (95% CI: 66.6, 72.4). A concordance probability significantly greater than 50% is indicative of a good predictive model.

17.4.4 External Validation

External validation was conducted to measure how well the Cox model derived from our registry works for the FIRST study described in Sect. 17.3.4. Ten variables which were significant predictors in the Cox model developed from the registry data were collected in the FIRST study in an identical manner. The variable "# of novel therapies" had levels defined as ≥2 novel therapies or (0, 1) novel therapies as part of the induction regimen in first line. Novel therapies being administered in cohort 1 of the registry included the multiple myeloma drugs Revlimid, Pomalidomide, Velcade, and Carfilzomib. In the FIRST study, patients were randomized to Revlimid + Dexamethasone continuous, Revlimid + Dexamethasone for 18 months and Melphalan, Prednisone, and Thalidomide for 18 months. The first of the three groups were most efficacious (Benboubker and Dimopoulos 2014) and was mapped to the ≥2 level and the remaining groups to (0, 1) of the novel therapy variable. These 11 variables as well as the survival duration and censoring variables were extracted from the FIRST database. Then the probability of survival beyond 3 years was computed for FIRST data using the registry-derived Cox model and compared against actual outcomes in the FIRST study. To compute the probability of survival beyond 3 years, we employ SAS code like that in Sect. 17.4.2 using the actual predictor combinations found in the FIRST study instead of the covals dataset. To compare actual outcomes in FIRST to predicted outcomes based on the registry outcomes, we use the following R code

```
library(rms)
surv.obj2 = with(dmm020cox,Surv(time, cens))
w <- rcorr.cens(x=dmm020cox$s,S=surv.obj2)
C  <- w['C Index']
se <- w['S.D.']/2
low <- C-1.96*se; hi <- C+1.96*se

library(rms)
S <- Surv(dmm020cox$time, dmm020cox$cens)
if('polspline' %in% row.names(installed.packages())) {
w <- val.surv(est.surv=dmm020cox$s, S=S, u=1095,
fun=function(p)log(-log(p)))
plot(w, xlab="Predicted Probability of Surviving Beyond 3 Years",
ylab="Actual Probability of Surviving Beyond 3 Years",
lim=c(.05,1),scat1d.opts=list(nhistSpike=200, side=1))
groupkm(dmm020cox$s, S, m=100, u=1095, pl=TRUE, add=TRUE)
text(0.4,0.95, "C-Index = 0.6787", cex = 0.9)
}
```

The first part of the code computes the concordance index and 95% CI as 67.8% (66.1, 69.6). This second part of the code produces Fig. 17.5.

The concordance probability compares favorably to 69.5% in the internal validation supporting the portability of the derived model. This is despite differences between FIRST, a CCT, and the registry. The dots in the graphic represent sets of about 100 patients each. The data plots above the equiangular line, likely, because FIRST had a healthier cohort due to restrictive inclusion and exclusion criteria. This

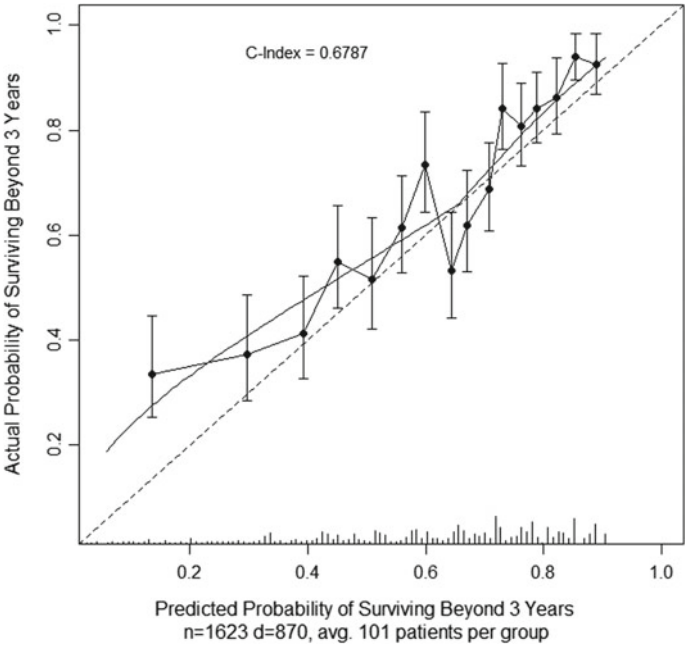


Fig. 17.5 External validation of the Cox model

would have lead to the higher actual then predicted probabilities of survival beyond 3 years.

17.5 Discussion

For a prognostic model for a disease to be meaningful, it should be based on data which is representative of the larger population of all patients with the disease, rather than a heavily screened cohort. Registries can provide such data as they have less restrictive inclusion and exclusion criteria, especially compared to controlled studies. However, registries tend to have more missing data due to less rigor in monitoring sites and leeway given to sites on procedures and observations to be recorded. Conventional model-building methods work for complete data and need to be modified considerably in the context of missing data. We have implemented model building for our incomplete registry dataset using analysis after multiple imputation. Such analysis is valid under missing at random conditions. This assumption implies that on using all information collected in the study, there remains no extraneous additional information predictive of missing values. An assumption which is likely to hold when a large amount of relevant information is collected and then used in the imputation model.

We describe the combination of inferences from each of the imputed datasets, modeling in the discrete and survival contexts, internal and external validation, and the construction of an intuitive representation of the prediction data. Through bootstrap re-sampling, we used the gold standard for internal validation to validate the prediction models. External validation described helps with the demonstration of the portability of the model to different patient populations. Similar model representations have been published in peer-reviewed journals (Vastesaeger et al. 2011; Vastesaeger et al. 2009) but a complete description of the methodology surrounding the modeling and construction of the matrix representation is currently unavailable, especially in the context of missing data. This chapter attempts to fill that void. The heat-map prediction matrix is a striking visual as well as numerical representation of prediction models.

For the selected illustrative examples, the models were externally validated using a large clinical trial. The concordance probabilities, measure of the predictive ability of the models, were 74.3 and 69.5% in the logistic and Cox internal validations and 71.8 and 67.8% in the external validations. These are in the ballpark of the concordance probability of 72.2% by internal validation for the Framingham heart study model, a Cox regression model (Pencina and D'Agostino 2004).

Acknowledgements The authors would like to thank Arlene Swern for her leadership and support during the development of the models described here and her review and edits of drafts of this document. The encouragement and feedback from the Connect® MM Registry study team and study steering committee members is much appreciated.

Appendix 1: SAS Code to Generate Predictions for Insertion into the Prediction Matrix

```
*****************************************************************;
**********Creation of Coval dataset*****************************;
*****************************************************************;
options orientation=portrait;
%let sitevar=8; *Number of blocks;
%let ptsvar=64; *number of cells;
%let blocksize=8; *number of cells per block;
%let ptspersite=%sysevalf(&ptsvar/&sitevar);
%let blockspersite=%sysevalf(&ptspersite/&blocksize,ceil);

%put &ptspersite;

proc plan ;
  factors block=&sitevar ordered pt=&blocksize ordered /noprint;
  output out=rsched;
run;
proc sort data=rsched;by block;run;
data rsched1;;
  set rsched;
   mobf=0; *Can be changed to 0 1 or 2;
   *mhecogynn=0; *Can be changed to 0 or 1;
  cell + 1;
  if first._n_ then cell = 1;
run;
data rsched1;
   set rsched1;
     if block in (1 2 3 4) then mhecogynn=1;
     if block in (5 6 7 8) then mhecogynn=0;
        if block in (1 2 5 6) then mhhyn=1;
        if block in (3 4 7 8) then mhhyn=0;
        if block in (1 3 5 7) then agen=1;
        if block in (2 4 6 8) then agen=0;
```

```
      if pt in (1 2 3 4) then plat_ct=0;
      if pt in (5 6 7 8) then plat_ct=1;
         if pt in (1 2 5 6) then issstagen=0;
         if pt in (3 4 7 8) then issstagen=1;
         if pt in (1 3 5 7) then creat=0;
         if pt in (2 4 6 8) then creat=1;
 run;

 data covals;
 set rsched1;
 drop pt block;
run;
proc sort data=covals;by cell;run;

data est2;
 set est1;
 keep parm estimate;
run;

*****************Transpose estimate**********************;
proc transpose data=est2 out=est3;
  id parm;
run;
*Macro to rename variable so that we can merge the transposed data
set with the main dataset covals;
%macro rename1(oldvarlist, newvarlist);
  %let k=1;
  %let old = %scan(&oldvarlist, &k);
  %let new = %scan(&newvarlist, &k);
     %do %while(("&old" NE "") & ("&new" NE ""));
      rename &old = &new;
         %let k = %eval(&k + 1);
      %let old = %scan(&oldvarlist, &k);
      %let new = %scan(&newvarlist, &k);
  %end;
%mend;

data est3;
 set est3;
 drop _NAME_ ;
 %rename1(agen issstagen mhecogynn mhhyn plat_ct creat mobf, agenm
issstagem mhecogynm mhhynm plat_ctm creatm mobfm);
run;

data est3b;
set est3;
do cell=1 to 64; /*Change to ptsvar number above in this case
ptsvar=32*/
       output;
       end;
run;

proc sort data=est3b; by cell; run;
proc sort data=covals; by cell; run;
data covals1;
 merge covals est3b;
 by cell;
run;
```

```
proc sort data=covals1;by cell; run;

*To sum Beta for each cell;
data covals2;
 set covals1;
 Intercept1=intercept;
 agens=agenm*agen;
 issstages=issstagem*issstagen;
 mhecogyns=mhecogynm*mhecogynn;
 *mhdiabns=mhdiabnm*mhdiabn;
 mhhyns=mhhynm*mhhyn;
 creats=creatm*creat;
 *mhsolynns=mhsolynnm*mhsolynn;
 *calciums=calciumm1*calcium;
 plat_cts=plat_ctm*plat_ct;
 mobfs=mobfm*mobf;
 *agdiab=agenmhdiabnm1*agenmhdiabn;
 *pltmob=plat_ctmobfm1*plat_ctmobf;
 sumbeta=agens+issstages+mhecogyns+mhhyns+plat_cts+creats+mobfs;
run;
*To calculate the Predicted Probability for each cell;
data pred;
 set covals2;
 x=Intercept1 + sumbeta;
 xy=1+exp(-x);
 prob=1/xy;
run;
proc sort data=pred;by cell;run;
proc sort data=rsched1;by cell;run;

data all;
 merge pred rsched1;
 by cell;
 keep cell agen issstagen mhecogynn plat_ct mhhyn mobf creat pt block
prob;
 format prob  percent10.;
run;

proc print data=all;run;
proc print data=all;var cell prob;run;
```

References

Allison, Paul D. (2010). *Survival analysis using SAS: A practical guide* (2nd ed.). Cary, NC: SAS Institute Inc.

Benboubker, L., Dimopoulos, M. A., et al. (2014). Lenalidomide and dexamethasone in transplant-ineligible patients with myeloma. *The New England Journal of Medicine, 371*(10), 906–917.

Connect® MM—The Multiple Myeloma Disease Registry. http://clinicaltrials.gov/ct2/show/NCT01081028.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology, 143,* 29–36.

Harrell, Frank E. (2001). *Regression modeling strategies with applications to linear models, logistic regression and survival analysis*. New York: Springer.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, New Jersey: Wiley.

Patricia, Bergland, & Steven, Heeringa. (2014). *Multiple imputation of missing data using SAS*. Cary, NC: SAS Institute Inc.

Pencina, M. J., & D'Agostino, R. B. (2004). Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Statistics in Medicine, 23*, 2109–2123.

Rifkin, R. M., Abonour, R., Terebelo, H., et al. (2015). Connect MM registry: The importance of establishing baseline disease characteristics. *Clinical Lymphoma Myeloma and Leukemia, 15,* 368–376.

Shah, J. J., Abonour, R., Gasparetto, C., Hardin, J. W., Toomey, K., et al. (2017). Analysis of common eligibility criteria of randomized controlled trials in newly diagnosed multiple myeloma patients and extrapolating outcomes. *Clinical Lymphoma, Myeloma and Leukemia*. https://doi.org/10.1016/j.clml.2017.06.013.

Stokes, M. E., Davis, C. S., & Koch, G. G. (2012). *Categorical data analysis using SAS* (3rd ed.). Cary, NC: SAS Institute Inc.

Terebelo, H., Srinivasan, S., Narang, M., Abonour, R., Gasparetto, C., Toomey, K., et al. (2017). Recognition of early mortality in multiple myeloma by a prediction matrix. *American Journal of Hematology*. https://doi.org/10.1002/ajh.24796.

Townsley, C. A., Selby, R., & Siu, L. L. (2005). Systematic review of barriers to the recruitment of older patients with cancer onto clinical trials. *Journal of Clinical Oncology, 23,* 3112–3124.

Unger, J. M., Barlow, W. E., Martin, D. P., et al. (2014). Comparison of survival outcomes among cancer patients treated in and out of clinical trials. *Journal of the National Cancer Institute, 106*(3).

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16,* 219–242.

Vastesaeger, N., Xu, S., Aletaha, D., St Clair, W., & Smolen, J. S. (2009). A pilot risk model for the prediction of rapid radiographic progression in rheumatoid arthritis. *Rhematology, 48,* 1114–1121.

Vastesaeger, N., van der Heijde, D., Inman, R. D., et al. (2011). Predicting the outcome of ankylosing spondylitis therapy. *Annals of Rheumatic Diseases*. https://doi.org/10.1136/ard.2010.147744.

Vist, G. E., Bryant, D., Somerville, L., et al. (2008). Outcomes of patients who participate in randomized controlled trials compared to similar patients receiving similar interventions who do not participate (Review). *Cochrane Database Systematic Reviews, 3*, MR000009.

Wood, A. M., White, I. R., & Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine, 27,* 3227–3246.

Yuan, Y. C. (2000). Multiple imputation for missing data: Concepts and new development. In *SUGI 25 Proceedings*, P267-25.

Chapter 18
Tepee Plots, Graphics for Structured Tabular Data, with Biopharmaceutical Examples

Shankar S. Srinivasan, Vatsala Karwe and Li Hua Yue

18.1 Introduction

Graphics for multi-dimensional data feature prominently in (Tufte 2001; Murrell 2011). The tepee plot, to be described here, adds an important tool to this collection (1:US Patent 7,495,673 B1. Filed June 4 2005, accepted 24 Feb. 2009). Free licenses for this graphic are available for research and academia at (http://www.resourcetepee.com/). A tepee maps structured tabular data—any data in rows and columns. The structure is an ordering of the rows of the tabular data, which often occurs naturally in data sets. When it does not exist naturally, an ordering can usually be created by the user. Some shuffling, ordering, or reorganization of the columns of the table can also enhance the graphic. The rows represent levels and the columns are attributes whose levels are of interest. The graphic is an easy to grasp depiction, compared to raw data and its mathematical transformations, of the distribution of the elements of a structured tabular data set along column attributes and row levels. Colors associated with the column attributes help identify dominant attributes.

We start our explication of the tepee plot tool with some pharmaceutical organizational structure examples. The methodology of constructing a tepee plot is explained through an example of a tepee plot illustrating the flare and containment of the Ebola epidemic in West Africa. We return to further examples after this section with biopharmaceutical operational and scientific applications. The manuscript concludes with a brief discussion.

S. S. Srinivasan (✉) · V. Karwe · L. H. Yue
Department of Biostatistics, Celgene Corporation, Summit, NJ, USA
e-mail: shsrinivasan@celgene.com

V. Karwe
e-mail: vatsalakarwe@gmail.com

L. H. Yue
e-mail: lyue@celgene.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_18

18.2 Pharmaceutical Organizational Structure Examples

Early application of this tool was in characterizing organizational structure. The popular human resource structure tool, an organizational chart, provides an overview of manpower distribution, but it does not provide a visualization of other resources utilized. For instance, the manpower-related payroll expenses are not reflected in an organizational chart. An organizational chart does not tie organizational structure to resource utilization unlike the tepee plot. The span of an organizational chart is limited by the smallest possible size of its blocks and their contents. This is less of a limitation for the tepee plot. What might take 50–60 pages in an organizational chart can easily be depicted in one page through this graphic. In applications to organizational structure, we refer to the structured table as a resource utilization matrix and the tepee plot as a resource tepee plot. This resource utilization at various functions and at differing levels within the organization is mapped onto a color-coded chart. This depiction of resource utilization is intended as a compelling alternative to the organizational chart, having more functionality. It has a color on the chart as dominant when the function associated with the color expends resources to a larger extent. The tepee plot provides a fair assessment of resource utilization across organizational functions and hierarchical levels; should this tool be used as an aid in growth, attrition, and reorganization decisions. Strong consideration of intangibles such as work-flows and individual talent, experience and socioeconomic factors should precede the use of aggregate insights of organizational structure tepees.

18.2.1 Tepee Plot for Full-Time Equivalent (FTE) Data

The starting point for constructing the graphic is the structured tabular matrix which we will call a resource utilization matrix in the context of organizational structure tepee plots. A hypothetical resource utilization matrix for FTE is provided in Table 18.1 for a typical biostatistics and programming department in a large pharmaceutical company. The vice president (VP) has been arbitrarily assigned to the early phase statistics group. We could alternatively have assigned 1/8 FTE to each of the functional groups reporting to him. VP + 1 are the VP's direct reports. There is one corresponding to each functional group. There are five personnel in the VP + 2 reporting level in the clinical statistics group, four at the VP + 2 level in the early phase statistical group, etc. Such a data set is easy to pull out of an organizational chart.

Figure 18.1 is the resource tepee plot corresponding to the data in Table 18.1. The length of each horizontal line represents the total number of personnel at each level of the biostatistics and programming department. The width of each individual section of this line corresponds to the total within a function at the level being considered. The figure has a scale identifying a line length corresponding to 40 employees. For instance, one can see the 137 employees in the 'VP + 3' level of the organization as the

Table 18.1 Resource utilization matrix for a hypothetical biopharmaceutical biostatistics and programming department

	Clinical statistics	Early phase statistics	EU statistics and programming	Non-clinical statistics	Medical affairs statistics	Health outcomes	Clinical programming	Scientific liaisons	Row totals
Vice president (VP)	0	1	0	0	0	0	0	0	1
VP+1	1	1	1	1	1	1	1	1	8
VP+2	5	4	7	5	8	4	6	3	42
VP+3	16	14	4	20	37	10	36	0	137
VP+4	57	41	0	34	0	0	88	0	220
VP+5	0	0	0	10	0	0	0	0	10
Col. totals	79	61	12	70	46	15	131	4	418

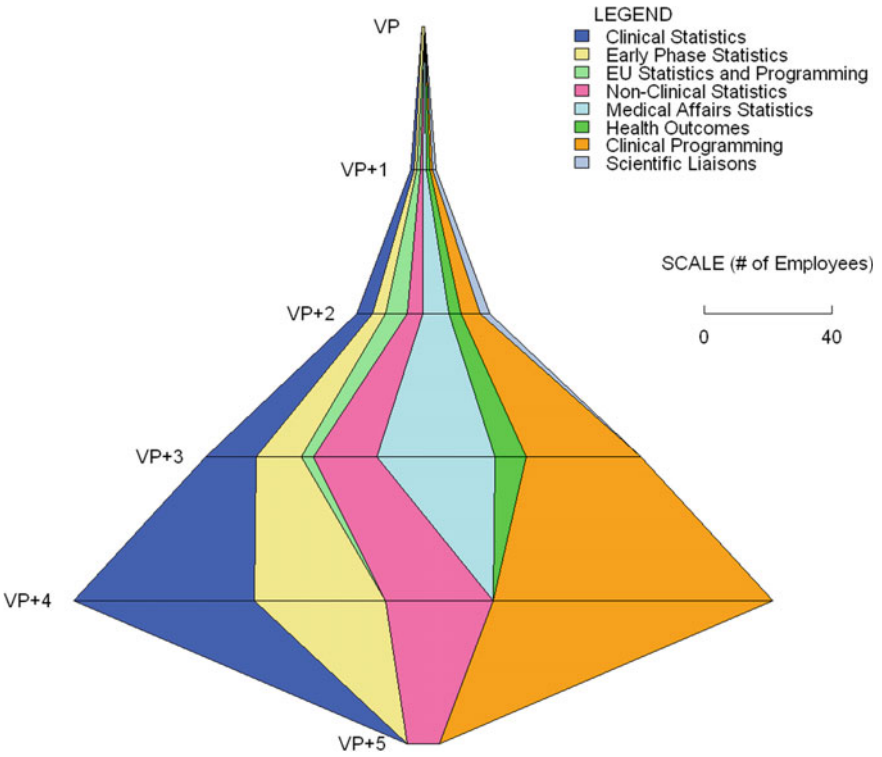


Fig. 18.1 Resource tepee plot for a biostatistics and programming department

entire length of the line at that level and the 16 employees in clinical statistics as the blue stretch of the line. Each level of the department is equidistant from the previous and the next level. The area corresponding to each function is invariant to reordering. To construct the tepee plot, the line corresponding to each row total is centered on the page and we keep adding segments for each function. Each color-shaded area is roughly proportional to the number of employees in the function representing the color.

18.2.2 Distribution of Personnel in a Simulated Pharmaceutical Company

In this example, we develop a visualization of the organizational structure for a simulated corporation. The simulation is based roughly on industry averages for the pharmaceutical and medicines industry as reported by the Bureau of Labor Statistics (http://www.bls.gov/). For instance, you will see in Table 18.2, 30% of the personnel

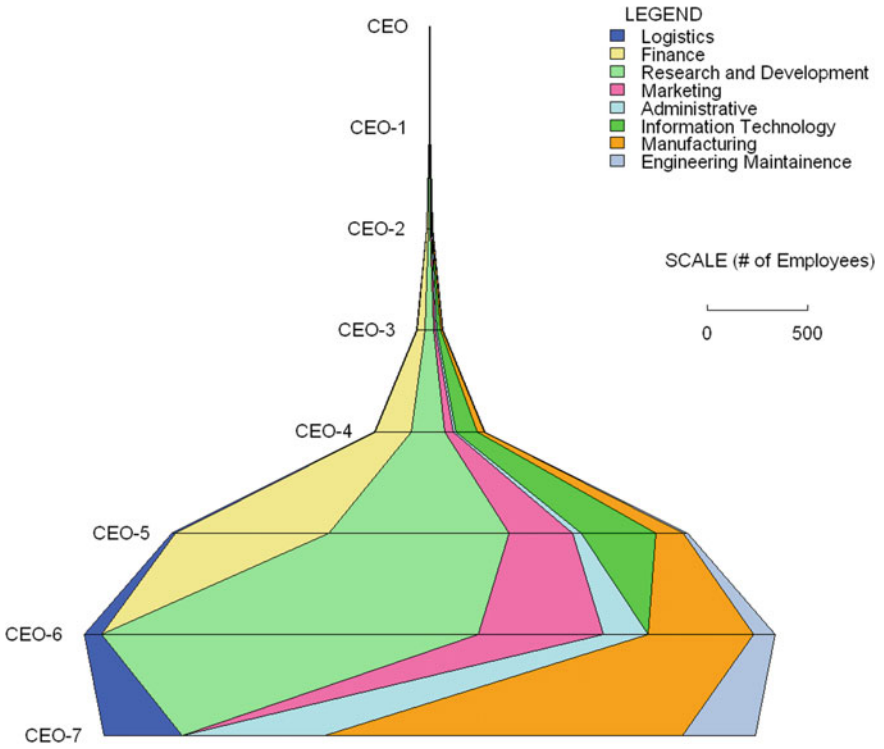


Fig. 18.2 Resource tepee plot for a simulated biopharmaceutical company

are in R&D and 25% are in manufacturing. These are based on the 31.4% reported in professional and related occupations and the 27.5% reported in production occupations in 2008 by the Bureau of Labor Statistics. However, we acknowledge that what is true for the industry may not necessarily be true for individual corporations. Further, the cascading of these employees through different levels or salary grade/bands in the organization are based on our estimates. A roughly four-to-five-fold increase in personnel is assumed when we go down a grade. A 10,000-employee corporation was simulated. Eight grade levels are assumed starting with the CEO and going down to CEO-7. The distribution to be displayed is in Table 18.2.

The visualization of this resource utilization matrix in Table 18.2 is in Fig. 18.2 . The organization is most numerous at the CEO-6 level. The scale adjacent to the figure identifies 500 employees. Research and development, represented by the green-colored segment, has 1880 employees at the CEO-6 organizational level. For the same simulated employee numbers, as in the examples in this section, the following section will look at compensation at various levels with the objective of looking at executive compensation.

Table 18.2 Resource utilization matrix for a simulated biopharmaceutical company

Grade	Logistics	Finance	R&D	Marketing	Administrative	IT	Manufacturing	Engg/maint	Row sum
CEO	0	0	0	1	0	0	0	0	1
CEO-1	0	2	2	1	1	0	2	0	8
CEO-2	0	10	10	4	2	3	3	1	33
CEO-3	1	40	42	10	6	20	8	2	129
CEO-4	4	180	168	44	12	106	33	6	553
CEO-5	18	768	898	316	43	371	140	21	2575
CEO-6	87	0	1880	624	223	0	528	107	3449
CEO-7	390	0	0	0	713	0	1786	363	3252
Column sum	500	1000	3000	1000	1000	500	2500	500	10000

18.2.3 Juxtaposition of CEO and Worker Compensation

The scrutiny of executive compensation has increased with the enactment of the Dodd-Frank Wall Street Reform and Consumer Protection Act (http://www.cftc.gov/LawRegulation/DoddFrankAct/index.htm). Shareholders can vote on executive compensation packages. While such a vote is non-binding, it can result in considerable negative publicity for a corporation and could lead to shareholder lawsuits as has been the case for Citigroup (Solomon 2012). The Resource Tepee can be useful as a tool to assess if executive compensation is excessive. A juxtaposition of CEO compensation with those across all employees at each of the other band levels (or within functions at chosen band levels) in an organization through this visualization may be helpful. We return to the simulated corporation in our example in Sect. 18.2.2 and look at the compensation spread across functions and levels. This simulation of compensation assumes a 350-fold increment of CEO compensation over that of the average worker in the lowest band of the corporation. Professor G William Domhoff in his Web site (Domhoff et al. 2016) on wealth, income, and power in America cites an average 344-fold increment in CEO compensation over that of an average worker in 2007. Compensation in multiples of 30, 11, 6, 4, 2.5, and 1.5 of the base worker is assumed as we go down the band levels. The Bureau of Labor Statistics reports a wage of $13.36 and $15.31 per hour in 2008 for two occupations at the lower end of the pharmaceutical and medicine industry's occupation categories. This simulation uses a compensation of $15.00 per hour at the lowest band. This average income drop-off is depicted in the graphic in Fig. 18.3 .

The wine glass-shaped graphic depicts a likely drop-off in average compensation as we go down the salary grades in an organization. However, we do expect the leaders of our industry to be reasonably well paid as their skills and talents can generate a lot of shareholder and employee wealth and they can make important contributions to our society. So, it is also fair to see an executive's compensation in the context of aggregate compensation across different levels and functions of the organization he heads. The resulting aggregate compensation resource utilization matrix can be obtained by multiplying the average compensation by the employee tallies in Table 18.2 and is not provided here.

The Resource Tepee in Fig. 18.4 is a visualization of the distribution of compensation. What constitutes excessive compensation is difficult to address. One may need a panel of sociologists, shareholder representatives, economists, and other involved coalitions to address compensation. Graphics can be an aid in making these decisions. One could get some insights on compensation by laying this tepee next to a tepee done for the distribution of personnel as well as looking at the wine glass-shaped average income drop-off down the hierarchy of an organization.

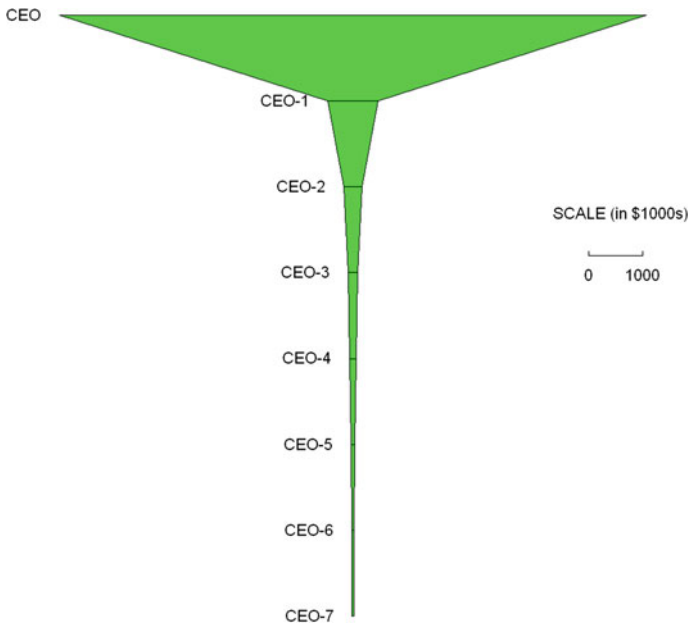


Fig. 18.3 Drop-off in average compensation through salary grades

18.3 Methodology

The tepee plot is a color display depicting, at a glance, the distribution of the elements of a structured tabular data across column attributes and row levels. This section will describe how this display is constructed by starting with this structured tabular matrix and computing a matrix of coordinates. This matrix contains the nodes of the tepee plot. Boundary lines to separate the colors of the plot are obtained by connecting coordinates in each column of the matrix of coordinates. Colors representing distinct attributes are applied to the spaces between boundary lines to obtain the tepee plot. Data on the Ebola epidemic is interspersed within the detailed description that follows to illustrate the construction of the tepee plot.

18.3.1 West African Ebola Data

The Ebola virus and its possible spread to the USA and elsewhere through air travel was a major concern during the 2014 outbreak. The primary locus of the epidemic was in the West African countries of Guinea, Liberia, and Sierra Leone. The US Center for Disease Control (CDC) maintained a tally of the total number of cases of the disease in these three countries at their Web site (https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/). Table 18.3 provides the number of new cases of

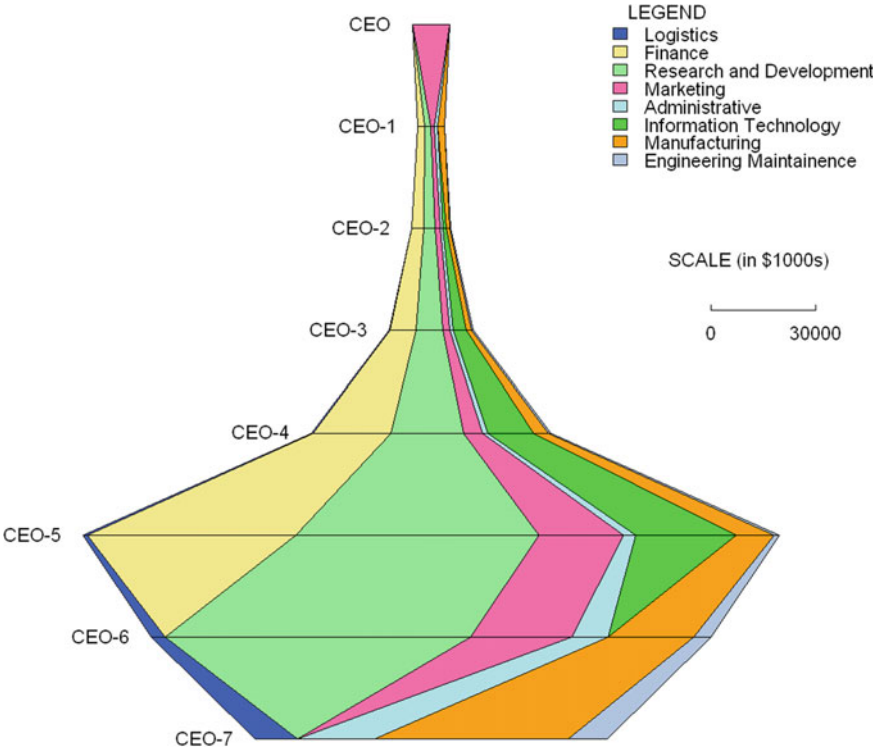


Fig. 18.4 Compensation tepee for a simulated corporation

Ebola in the three West African countries as inferred from the CDC data. The disease flared rapidly from April to October 2014 and it took about a year to contain it.

Table 18.3 West African Ebola data

Month	New cases monthly, Guinea	New cases monthly, Liberia	New cases monthly, Sierra Leone	Row sum
OCT 2015	3	0	156	159
SEP 2015	15	0	308	323
AUG 2015	9	0	216	225
JUL 2015	36	6	258	300
JUN 2015	93	0	302	395
MAY 2015	74	344	429	847
APR 2015	86	610	424	1120
MAR 2015	337	474	673	1484
FEB 2015	238	616	783	1637

(continued)

Table 18.3 (continued)

Month	New cases monthly, Guinea	New cases monthly, Liberia	New cases monthly, Sierra Leone	Row sum
JAN 2015	238	604	1072	1914
DEC 2014	552	383	2337	3272
NOV 2014	488	1100	1771	3359
OCT 2014	593	3077	3317	6987
SEP 2014	426	2080	995	3501
AUG 2014	188	1049	493	1730
JUL 2014	70	278	375	723
JUN 2014	109	39	142	290
MAY 2014	60	0	16	76
APR 2014	109	4	0	113
COL SUM	3724	10664	14067	28455

18.3.2 Structured Tabular Matrix

The structured tabular matrix for the Ebola data follows:

$$L = \begin{bmatrix} 3 & 0 & 156 \\ 15 & 0 & 308 \\ 9 & 0 & 216 \\ 36 & 6 & 258 \\ 93 & 0 & 302 \\ 74 & 344 & 429 \\ 86 & 610 & 424 \\ 337 & 474 & 673 \\ 238 & 616 & 783 \\ 238 & 604 & 1072 \\ 552 & 383 & 2337 \\ 488 & 1100 & 1771 \\ 593 & 3077 & 3317 \\ 426 & 2080 & 995 \\ 188 & 1049 & 493 \\ 70 & 278 & 375 \\ 109 & 39 & 142 \\ 60 & 0 & 16 \\ 109 & 4 & 0 \end{bmatrix}$$

Essential notation is that the structured tabular matrix is denoted by $\boldsymbol{L}$, consisting of elements l_{kj} which represents the numerical quantity to be displayed for the kth attribute at the jth level. The subscript k for the columns from the left to the right of the matrix varies from 1 for the first attribute to the last or nth attribute in some appropriate preselected order. The subscript j for the rows from the bottom to the top of the matrix varies from 0 for the bottom most in the hierarchy to some highest level m.

To clarify the notation, note, for instance, that l_{23} is the Ebola tally for the second attribute at the third level in the hierarchy (from Table 18.3, it is the number of new cases in Liberia in July 2014). From the matrix l_{23} is 278. Similarly, l_{30} is the number of cases for the third attribute at the 0th or bottom level and equals 0, $l_{11} = 60$.

18.3.3 Matrix of Coordinates

The next step in the process computes a matrix $\boldsymbol{C}$ of coordinates. This matrix contains the elements c_{ij} given by

$$c_{ij} = \left\{ \left(\sum_{k \le i} l_{kj} - 0.5 * \sum_{k=1}^{n} l_{kj} \right), a * j \right\}$$

where $\sum$ is a symbol for a summation and a is some appropriate amount by which we would like the hierarchical levels in the tepee to be separated. A comma separates the x coordinate from the y coordinate. The subscript i for the columns from left to right of the matrix $\boldsymbol{C}$ varies from 0 to n and represents the n + 1 boundaries which will separate the colors of the tepee. The subscript j for the rows from bottom to top of the matrix $\boldsymbol{C}$ varies as before from 0 for the bottom most to the highest level m. The l_{kj} are the elements of the resource utilization matrix as defined earlier. The coordinates in $\boldsymbol{C}$ are the nodes of the tepee.

To clarify the computation of the $\boldsymbol{C}$ matrix, let us return to our Ebola example. For the resource utilization matrix in our example, the matrix $\boldsymbol{C}$ of tepee coordinates computes as shown in the following page.

Let's verify c_{21}. Using the formula, this coordinate is given by

$$c_{21} = \left\{ \left(\sum_{k \le 2} l_{k1} - 0.5 * \sum_{k=1}^{3} l_{k1} \right), a * 1 \right\}$$

The quantity $\sum_{k \le 2} l_{k1}$ is the sum of the elements the first tier for the first two attributes. Using the $\boldsymbol{L}$ matrix, this is equal to $60 + 0 = 60$. The quantity $\sum_{k=1}^{3} l_{k1}$ is the sum over all attributes in the first tier. This equals $60 + 0 + 16 = 76$. Using $a = 1$

$$c_{21} = \{(60 - 0.5 * 76), 1\} = (22, 1)$$

Similarly consider

$$C_{33} = \left\{ \left(\sum_{k \le 3} l_{k3} - 0.5 * \sum_{k=1}^{3} l_{k3} \right), a * 3 \right\}$$
$$= \{(70 + 278 + 375 - 0.5 * (70 + 278 + 375)), 3\} = (361.5, 3)$$

Note that for i = 0, the first summation drops out and

$$c_{0j} = \left\{ -0.5 * \sum_{k=1}^{8} l_{kj}, a*j \right\}$$

$$C = \begin{bmatrix}
(-79.5, 18) & (-76.5, 18) & (-76.5, 18) & (79.5, 18) \\
(-161.5, 17) & (-146.5, 17) & (-146.5, 17) & (161.5, 17) \\
(-112.5, 16) & (-103.5, 16) & (-103.5, 16) & (112.5, 16) \\
(-150, 15) & (-114, 15) & (-108, 15) & (150, 15) \\
(-197.5, 14) & (-104.5, 14) & (-104.5, 14) & (197.5, 14) \\
(-423.5, 13) & (-349.5, 13) & (-5.5, 13) & (423.5, 13) \\
(-560, 12) & (-474, 12) & (136, 12) & (560, 12) \\
(-742, 11) & (-405, 11) & (69, 11) & (742, 11) \\
(818.5, 10) & (-580.5, 10) & (35.5, 10) & (818.5, 10) \\
(-957, 9) & (-719, 9) & (-115, 9) & (-957, 9) \\
(-1636, 8) & (-1084, 8) & (-701, 8) & (1636, 8) \\
(-1679.5, 7) & (-1191.5, 7) & (-91.5, 7) & (1679.5, 7) \\
(-3493.5, 6) & (-2900.5, 6) & (176.5, 6) & (3493.5, 6) \\
(-1750.5, 5) & (-1324.5, 5) & (755.5, 5) & (1750.5, 5) \\
(-865, 4) & (-677, 4) & (372, 4) & (865, 4) \\
(-361.5, 3) & (-291.5, 3) & (-13.5, 3) & (361.5, 3) \\
(-145, 2) & (-36, 2) & (3, 2) & (145, 2) \\
(-38, 1) & (22, 1) & (22, 1) & (38, 1) \\
(-56.5, 0) & (52.5, 0) & (56.5, 0) & (56.5, 0)
\end{bmatrix}$$

18.3.4 Constructing the Tepee Plot

The coordinates in the matrix ***C*** in our example are plotted in Fig. 18.5. The y-axis is a scale from 0 to 18 with tick marks 1 unit apart. Horizontal lines in the plot go through row coordinates, and the scale for the horizontal lines for 1000 Ebola cases is provided to the right-hand side. Once the nodes have been plotted, the boundaries

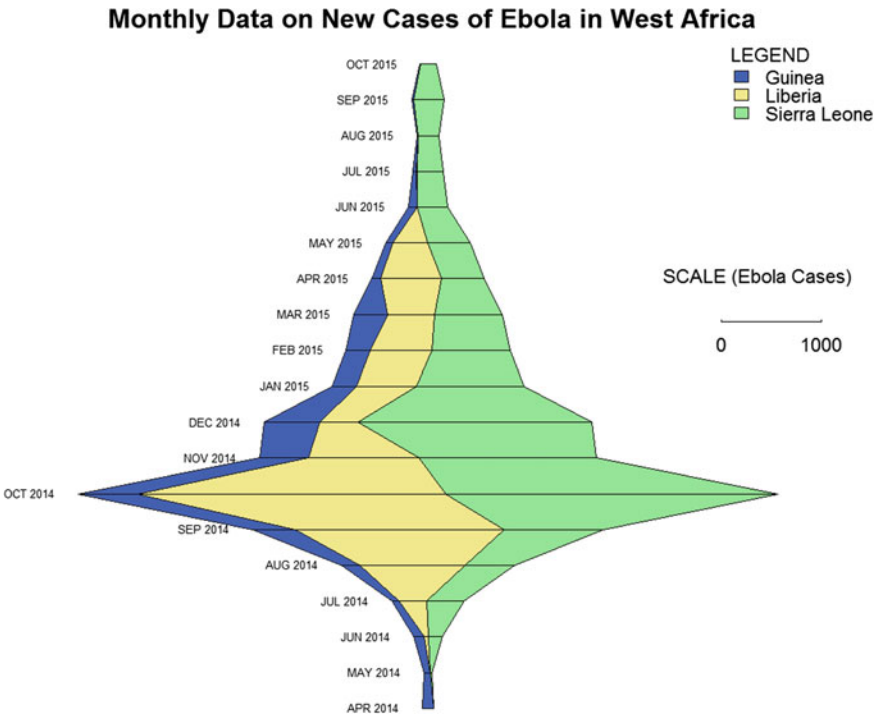


Fig. 18.5 Tepee plot for new cases of Ebola in West Africa

between colors are obtained by connecting the coordinates in each column of the matrix of coordinates. For instance, the outer left boundary of the tepee is obtained by tracing a line through the coordinates (−79.5, 18), (−161.5, 17), etc., down to (−38, 1) and (−56.5, 0) in the first column of $\boldsymbol{C}$. The second line from the left is obtained by connecting (−76.5,18), (−146.5, 1), etc., down to (22, 1) and (52.5, 0) in the second column of $\boldsymbol{C}$. Proceeding in this manner through all the columns of $\boldsymbol{C}$ in our example we obtain all the boundary lines in Fig. 18.5.

Once the boundary lines are obtained, we associate n distinct colors to the n attributes. Starting from the left, we fill in the space between the boundary lines with colors going from that for the first column attribute to that for the last or nth attribute. In our example, we chose blue for the first attribute, yellow for the second attribute, and green for the third attribute. Using these colors, the tepee in Fig. 18.5 is obtained.

18.4 Other Operational and Scientific Biopharmaceutical Applications

Tepee plots can be used in broader contexts outside organizational structure. There are many other contexts in which we find structured tabular data. Any table in rows and columns with any quantitative content and some natural or induced ordering can be displayed using a tepee plot. In this section, we look at tepee plots for outsourcing costs for clinical trials, laboratory grade shift data tepee plots, and tepee plots for cancer therapies over time.

For the outsourcing and grade shift tepee plots, we add null lower and upper nodes—all tepee plot boundaries are extended to a point with x coordinate = 0 at one level below and above. We had noted earlier, when discussing resource tepee plots that without adding such null nodes the total area covered by each color was roughly proportional to the total associated with the column function. This proportionality is exact when we add the null upper and lower nodes. This follows on noting that the area corresponding to a column attribute is given by the sum of the areas of the top and bottom triangles and the summation of the trapezoidal areas in between. This area computes to the following for every column j

$$0.5 * l_{0j} * a + 0.5 * l_{mj} * a + \sum_{i=0}^{m-1} 0.5 * \left(l_{ij} + l_{i=1,j}\right) * a = a * \sum_{i=0}^{m} l_{ij}$$

Thus, the area corresponding to a column attribute is proportional to the total associated with the column attribute with a proportionality constant given by a, the distance between the horizontal lines in the plot.

18.4.1 Outsourcing Cost Tepee

Pharmaceutical companies find it expedient to outsource some or all the work involved in running a clinical trial to CROs. This example looks at the distribution over different functions of costs in outsourcing budgets. In the other examples, there was a hierarchy, levels in an organization, which spanned the vertical space of the graphic. Through this example, we illustrate that such organizational hierarchies are not necessary, and as noted elsewhere, any quantitative measure which can be represented in a two-way data matrix can be represented by the resource tepee. The costs of outsourcing are typically associated with the number of patients (N) that are to be enrolled in the study (though this association can be thrown off somewhat by special needs in the studies such as the collection and analysis of biomarker data, special study design features, clinical phase of the study). In this example, the number of patients in the clinical studies being displayed is used to order the vertical space of the graphic. More generally, one could create an ordering by using the row sums in the resource utilization matrix. The resource utilization matrix in Table 18.4

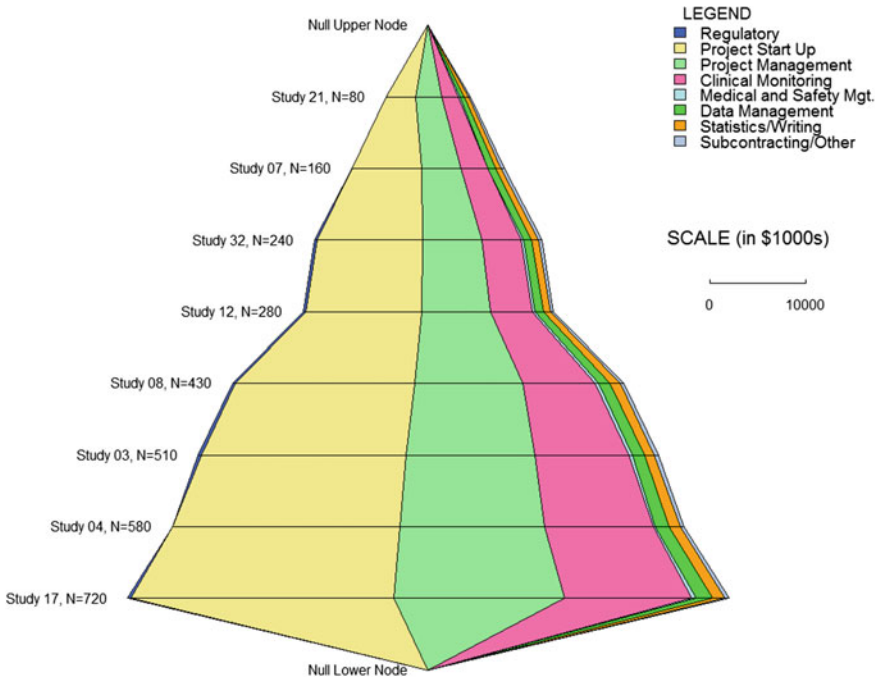


Fig. 18.6 Tepee plot for outsourcing costs of clinical trials

shows the distribution of outsourcing costs over a collection of clinical studies. The collection of studies could be defined as some meaningful collection such as one for those pertaining to a specific therapeutic area/drug or for a specific CRO. The idea is to develop a graphic which provides a holistic depiction of the distribution of costs in a slew of outsourced studies. The numbers in the resource utilization matrix that follows are made-up numbers (in $1000s) which bear no resemblance to any real outsourcing effort.

The graphic in Fig. 18.6 summarizes the data in Table 18.4 and adds a null upper node and a null lower node. These null nodes make the area of the colors in the graphic exactly proportional to the total cost, across studies, at the function associated with the color.

18.4.2 Laboratory Grade Shift Tepee

In clinical trials, laboratory values are graded in severity using some system such as the NCI CTCAE grading system for cancer drug trials (Common terminology criteria for adverse events 2016). Table 18.5 is a typical presentation of data about shifts in safety grades from the baseline to the worst grade while on treatment (this

Table 18.4 Outsourcing costs by clinical trial

	Regulatory	Project start-up	Project management	Clinical monitoring	Medical and safety mgt.	Data management	Statistics/writing	Subcontracting/other	Row sum
Study 21, N = 80	0	3124	2756	1700	190	479	417	204	8870
Study 07, N = 160	0	7234	4124	2876	0	769	589	317	15909
Study 32, N = 240	212	11031	6175	4100	287	879	669	354	23707
Study 12, N = 280	314	12073	7245	4301	315	912	715	287	26162
Study 08, N = 430	307	18756	11237	7568	324	1234	1079	371	40876
Study 03, N = 510	289	21478	13421	9861	354	1267	996	425	48091
Study 04, N = 580	0	23768	15123	11345	200	1450	1179	382	53447
Study 17, N = 720	355	27456	17843	13289	423	1765	1289	415	62835
Column sum	1477	124920	77924	55040	2093	8755	6933	2755	279897

Table 18.5 Shifts in anemia grades from baseline to worst grade on treatment

	Normal	Grade 1	Grade 2	Grade 3	Row sum
Baseline grade 3	0	14	20	0	34
Baseline grade 2	14	101	85	10	210
Baseline grade 1	40	162	45	7	254
Baseline normal	61	86	10	0	157
Column sum	115	363	160	17	655

table is created using made-up data which does not bear any resemblance to any real clinical trial). This is how you read the table:

- Thirty-four patients had a grade 3 anemia at baseline and that improved to a worst grade of 2 in 20 patients and grade 1 in 14 patients.
- One hundred and fifty-seven patients were normal on anemia at baseline. Sixty-one stayed normal on treatment, 86 had a worst grade of 1, and 10 had a worst grade of 2.

Figure 18.7 provides an intuitive visualization of this data. The length of each horizontal line represents the row total—the total number of patients in each baseline anemia grade. The segments of each horizontal line represent the total in this baseline grade who end up in the different worst grades on treatment as indicated by the colors in the legend. The number of patients with any combination of baseline grade and worst grade can be read off from the length of the associated segment using the scale provided. The use of the upper and lower null nodes ensures that the area of the colors is directly proportional to the column totals—the number of patients in each worst grade.

18.4.3 Cancer Therapies Over Time

The tepee plots in this section are those for multiple myeloma patients in second line in the time from 2009 to 2015 in the MM Connect Registry (see, Rifkin et al. 2015; http://clinicaltrials.gov/ct2/show/NCT01081 for details about the registry). These plots together with those for first-line induction and first-line maintenance were presented as a poster with additional clinical background and study details at the American Society for Hematology (ASH) 2016 conference at San Diego (Rifkin et al. 2009).

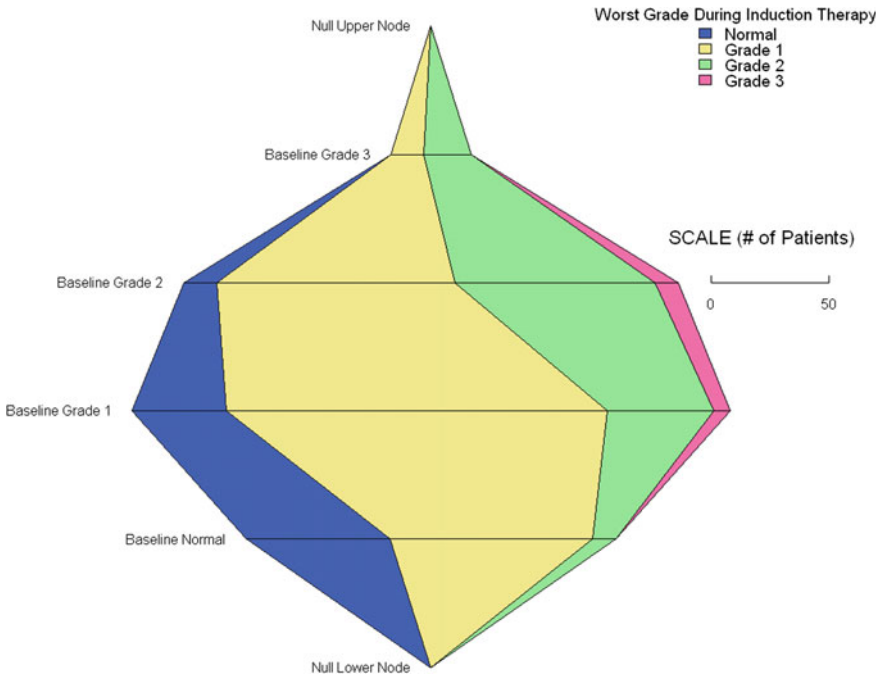


Fig. 18.7 Grade shift tepee plots

Each horizontal line in the first plot (Fig. 18.8) represents patients entering second-line induction/salvage during that period (read periods in the plots as in this example: Q34_2010 is the last two quarters of 2010). Each complete line represents 100% of the patients entering in that period. Using percentages with each line representing 100% gives these graphics a rectangular shape unlike other tepee plots presented in this document. We note that this special case rectangular tepee plot is somewhat like a graphic developed in the context of state transitions (Gabadinho et al. 2010). Each line is split into colored segments representing the percentages for each of the regimens in the legend. The regimens from left to right in the graphic map to the regimens from top to bottom in the legend.

A good number of patients in the graphic above fall in the gray 'other' category, and the Figure 18.9 visualizes the plethora of regimens given to patients in the US context. The total length of each horizontal line represents the total 'other' percentage in the corresponding period with segments for each regimen.

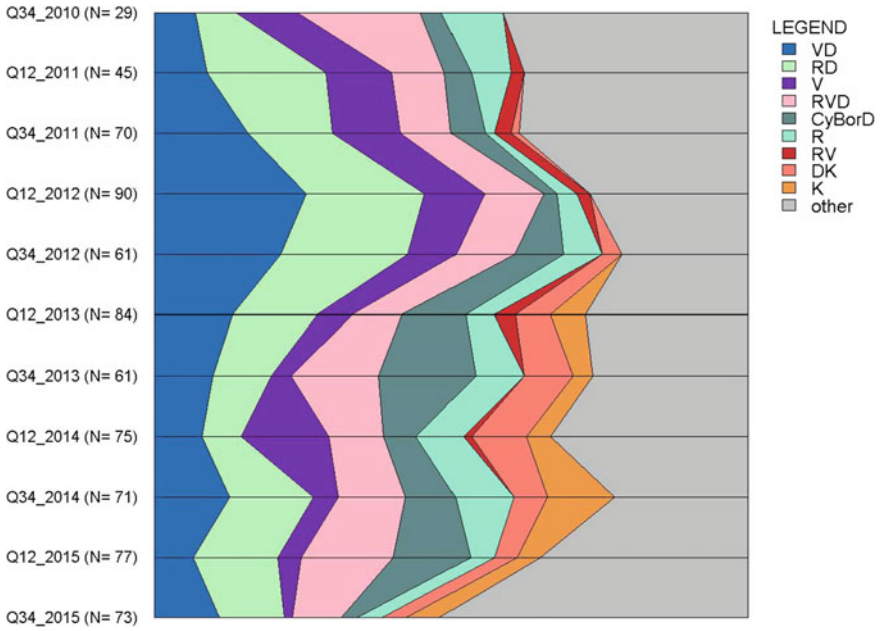


Fig. 18.8 Tepee plot of second-line cancer regimens over time

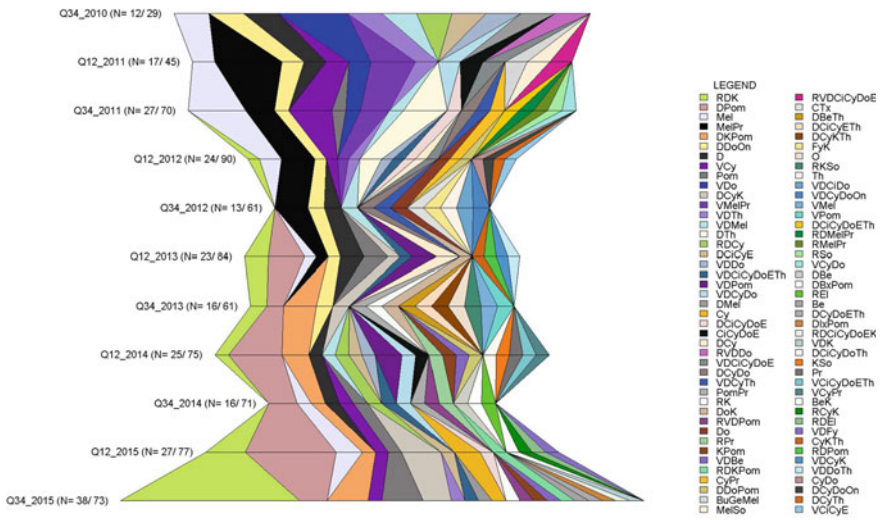


Fig. 18.9 Tepee plot with detail on other second-line cancer regimens over time

18.5 Discussion

The tepee plot can be constructed for any quantitative measure, available across functions/attributes and levels, in the form of a table with columns and rows. The tepee plot provides a one-page graphic to summarize structured tabular data, which can use an intuitive representation instead of the usual large spreadsheet tables. The tepee plot is easy to interpret. The extent and location of the colors tell us the extent to which an input of interest is distributed over column attributes and row levels. Many applications are possible as any table in rows and columns with any quantitative content and some natural or induced ordering can be displayed using this graphic.

For organizational structure tepees, resources plotted could include any quantifiable financial activity in an organization and professional groupings as well as groups of products, locations, and markets. Further, one could construct tepee plots for an organization or restrict the display to business units, processes, or projects.

The tepee plot makes three-dimensional data easy to read and understand. Many applications in research and practice as well as in organizational structure are anticipated.

Acknowledgements The authors would like to thank Arlene Swern for her leadership and support of the application of the tepee tool to pharmaceutical data and for her review and edits of drafts of this document. The encouragement and feedback from the Connect® MM Registry study team and study steering committee members are much appreciated.

References

2014–2016 Ebola Outbreak in West Africa. (2016). Retrieved December 07, 2016, from https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/.

Common terminology criteria for adverse events. (2016). Retrieved December 07, 2016, from http://www.uptodate.com/contents/common-terminology-criteria-for-adverse-events.

Connect® MM- The Multiple Myeloma Disease Registry. http://clinicaltrials.gov/ct2/show/NCT01081028.

Domhoff, W. G. Who Rules America: Wealth, Income, and Power. Retrieved December 07, 2016, from http://www2.ucsc.edu/whorulesamerica/power/wealth.html.

Dodd-Frank Act–CFTC. http://www.cftc.gov/LawRegulation/DoddFrankAct/index.htm.

Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. S. (2010). *Mining sequence data in R with the TraMineR package: A user's guide*. University of Geneva. http://mephisto.unige.ch/traminer.

Murrell, Paul. (2011). *R Graphics* (2nd ed.). Boca Raton, FL: CRC Press.

Rifkin, R. M., Abonour, R., Durie, B., Gasparetto, C. J., Jagannath, S., Narang, M., et al. Treatment patterns from 2009 to 2015 in patients with newly diagnosed multiple Myeloma in the United States: A report from the Connect® MM Registry. In *Submitted to the ASH 2016 Conference*, San Diego, California.

Rifkin, R. M., Abonour, R., Terebelo, H., et al. (2015). Connect MM registry: the importance of establishing baseline disease characteristics. *Clinical Lymphoma Myeloma and Leukemia, 15,* 368–376.

Solomon, S. D. (2012). Citigroup Has Few Options After Pay Vote. Retrieved December 07, 2016, from http://dealbook.nytimes.com/2012/04/18/citigroup-has-few-options-after-pay-vote/.

Srinivasan, S. (2009). Resource Tepee. Patent US 7,495,673 B1. February 24, 2009. Filed June 4, 2005.
Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.
Resource Tepee. http://www.resourcetepee.com/.
US Bureau of Labor Statistics. http://www.bls.gov/.

Chapter 19
Some Methods for Longitudinal and Cross-Sectional Visualization with Further Applications in the Context of Heat Maps

Shankar S. Srinivasan, Li Hua Yue, Rick Soong, Mia He, Sibabrata Banerjee and Stanley Kotey

19.1 Introduction

We will be presenting, in this chapter, a heuristic which builds on hierarchical clustering approaches and demonstrates its utility in ordering rows, typically representing subjects, in state sequence graphics, and, in ordering rows and columns, typically representing distinct samples and genes, in heat maps. We will first introduce graphics for longitudinal state sequence data as discussed by Gabadinho and colleagues (Gabadinho et al. 2010, 2011) with the intent of applying it to clinical data in cancer studies, to characterize the movement of patients over time through various response and disease states.

S. S. Srinivasan (✉) · L. H. Yue · S. Banerjee · S. Kotey
Department of Biostatistics, Celgene Corporation, Summit, NJ, USA
e-mail: shsrinivasan@celgene.com

L. H. Yue
e-mail: lyue@celgene.com

S. Banerjee
e-mail: sibanerjee@celgene.com

S. Kotey
e-mail: skotey@celgene.com

R. Soong · M. He
Department of Statistical Programming, Celgene Corporation, Summit, NJ, USA
e-mail: rsoong@celgene.com

M. He
e-mail: mhe@celgene.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_19

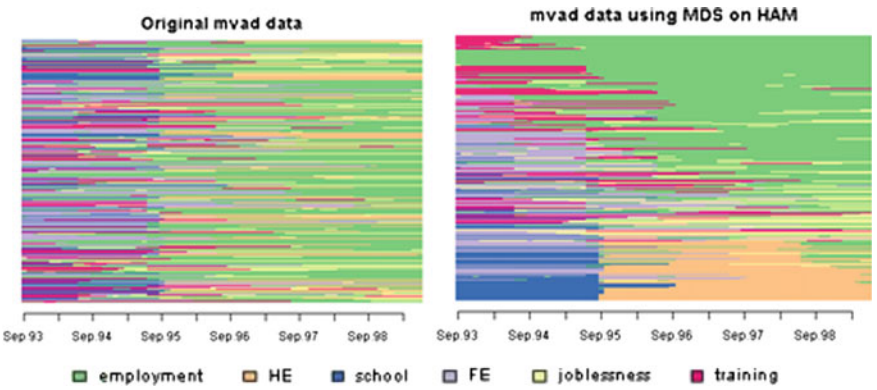


Fig. 19.1 Unordered employment and education sequences (left panel) and sequences ordered using multi-dimensional scaling on the Hamming distances between subjects (right panel)

19.1.1 Longitudinal State Sequence

Visualizations of sequences of subject states have been developed in response to the need for characterizing key social transitions over time as reported in (Giele and Elder 1998; Elder et al. 2003). The TraMineR R package (Gabadinho et al. 2016) comes with data collected in McVicar and Anyadike-Danes (2002) consisting of transitions between states characterizing education and employment in Northern Ireland from 1993 to 1999. This R package has considerable functionality, including descriptive and inferential analysis of sequences. We are particularly interested in depicting longitudinal patterns over subjects while preserving cross-sectional (sections by time) information. We offer an improvement in this context to plots generated using the methods in Gabadinho et al. (2010, 2011). Figure 19.1 shows unordered sequence data in the left panel and a display of data ordered using multi-dimensional scaling to the right. Each thin horizontal strip in the plots represents for each subject a sequence of states over time in the context of employment and education. The R Code follows

```
#call TraMineR R package and read the mvad data set
library(TraMineR)
data(mvad)
#create state sequence object and compute the HAM distance
measure between subject strings of transitions
mvad.alphab <- c("employment", "FE", "HE", "joblessness",
"school", "training")
mvad.seq <- seqdef(mvad, 17:86, xtstep=6,
alphabet=mvad.alphab)
HAMdist <- seqdist(mvad.seq, method="HAM")
#plot of unordered sequences
seqIplot(mvad.seq, sortv=1:712, cex.legend=0.9, ylab=NA,
yaxis=FALSE, title="Original mvad data")
#Ordering using multi-dimensional scaling
mds2 <- cmdscale(HAMdist,k=1)
seqIplot(mvad.seq, sortv=mds2, cex.legend=0.9, ylab=NA,
yaxis=FALSE, title="mvad data using MDS on HAM")
```

The graphic to the right in Fig. 19.1 orders the raw data to bring out some longitudinal as well as cross-sectional patterns in the data and uses an ordering based on the first dimension of a multi-dimensional scaling analysis derived from similarities computed using the Hamming method. Details on this method are in Sect. 19.2.1. In the right panel, we see subject similarities on the employment (green), higher education (orange) and school (blue) states, which are not apparent in the raw data. We seek to improve on this graphic through the edge clustering heuristic described in Sect. 19.2.5. Clinical application to subject transitions between responder categories over time, while on cancer therapy will be presented in Sect. 19.4.2. Our heuristic applies to the ordering of the columns as well, together with the rows—a brief introduction of this in the next section.

19.1.2 Two-Way Heat Maps

In two-way heat maps, typically for gene sample data, we have distinct genes as columns and distinct samples as rows. We have numeric data for every row column cell which is typically a scaled measure of gene expression. This is mapped monotonically to elements of a color palette changing gradually from one color and intensity at one end to another at the other end. The rows and columns are re-ordered to call out sets of genes and samples which have similar expression profiles. Figure 19.2 below is gene sample heat map based on gastric cancer data from The Cancer Genome Atlas (TCGA) Study (The Cancer Genome Atlas Research Network 2014) provided with the dendsort R package (Sakai 2015).

The data plotted in the heat map is the scaled association between genes (columns) and samples (rows). Row to row distance measures between every pair of rows are used to order the rows using the complete hierarchical clustering method (Sect. 19.2.3). The columns are ordered in a similar manner, and the row/column ordered data are plotted. In this figure, the scaled association measure is mapped to 15 color tones going from blue to red through white. A further sorting of the ordering obtained by hierarchical clustering is provided by Sakia et al. (2014) and is described in Sect. 19.2.4. We look briefly at existing methods, the edge clustering, and the framework for evaluating these methods in the next section.

19.1.3 Brief Overview of Ordering Methods and Their Evaluation

We develop our plots by starting with a distance or similarity measure between every pair of the rows and columns (Sect. 19.2.1). These measures are used to identify sets of rows providing similar patterns of transition for sequence data and arrangements of rows and columns for two-way heat maps presenting numeric data

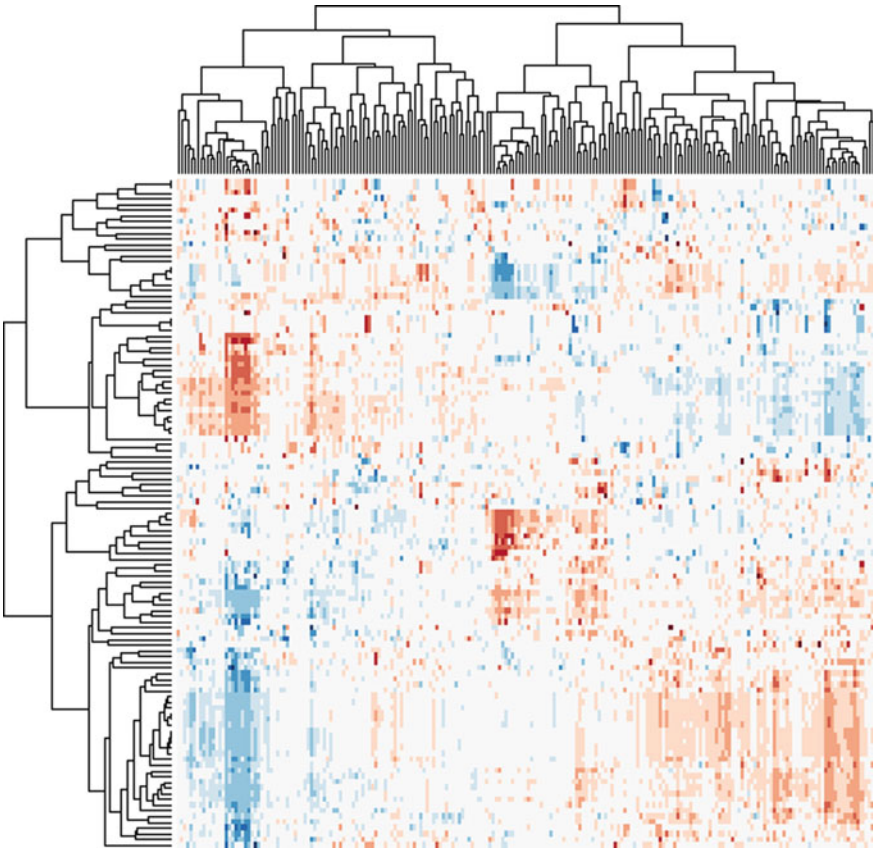


Fig. 19.2 Heat map of the scaled association of genes and samples using gastric cancer data from the TCGA study

as cohesive color terrains. Some methods which achieve this re-arrangement include multi-dimensional scaling, hierarchical clustering, sorted dendrograms, and the edge clustering method. The edge clustering method is developed within the framework of hierarchical clustering and will be tested using a conventional statistical evaluation framework. Such a framework often involves generation of data from a known non-standard distribution and an evaluation of the relative accuracy of various estimating processes, blinded to the parent distribution, in uncovering characteristics of the parent distribution. By analog, we will look at known latent images with row–column pixels containing color intensity data, randomly permute the rows and columns to lose all ordering information, and evaluate the ability of various methods to recover the latent image.

Table 19.1 Examples of operations that transform the string S1 to S2

S1	A	A	C	B	C		
S2	A	C	B	B	B		
3 Substitutions							
S1	A	A	C	B	C		
		S	S		S		
S2	A	C	B	B	B		
2 Insertions and 2 Deletions							
S1	A	A	C	B	C		
	D				D	I	I
S2?	–	A	C	B	–	B	B

19.2 Ordering Methods

We will look at applications where we order column or row strings of data consisting of discrete ordinal and nominal states, and continuous numeric data. Distance and similarity measures between these strings of data are utilized by ordering heuristics to assign them to appropriate rows in a sequence plot or appropriate rows and columns in a two-way heat map. Distance measures can be obtained from similarity measures by using any monotonic decreasing function. This section starts with such measures and then describes methods to order data strings. Technical details are provided for the edge heuristic. Other techniques are presented at an intuitive level. Any information deficit can be addressed by standard textbooks such as Anderson (1958) and Johnson and Wichern (1992).

19.2.1 Prerequisites: Distance and Similarity Measures

In the sequence data, introduced earlier, involving transition between education and employment states, it is hard to argue any ordering across the states. For such nominal data, several similarity measures are derived based on the number of operations which transform one sequence into another. Table 19.1 has an example from Gabadinho (2013), a source which describes these methods in detail. The operations considered are substitution, insertion, and deletion and typically have associated costs.

Popular measures include the Hamming distance (Hamming 1950) which involves substitutions alone. In Table 19.1, three substitutions (S) can get us from the string S1 to the string S2. A unit substitution cost would lead to a Hamming distance of 3. The optimal matching distance (Levenshtein 1966) allows substitutions, insertions (I), and deletions (D). The table obtains S2 from S1 using 2 deletions and 2 insertions. We input both the Hamming and the optimal distances into our ordering heuristics. In addition, we considered a pixel measure which is based on comparisons of an element in a string to the corresponding element of another, as well as the two

elements adjacent. This is based on the premise that in a visual display a pixel should ideally match with as many of the 8 pixels that surround it as possible. In our analyses, the ordering heuristics had more of an impact on the sequence plots than the choice of distance measures.

For strings which consist of continuous data, popular distance measures include Euclidean, Manhattan, maximum and Mahalanobis distances. The Euclidian is obtained as the vector norm of the differences between corresponding elements of two strings of numeric data. The Manhattan difference is the sum of absolute position-wise differences between two strings. The maximum distance is the maximum of such absolute differences.

We will be looking at applying sequence transition plots for discrete states reflecting response levels, the state of having progressed, and death. These states can be rank ordered. With this ordinal data, we will use the same distance measures as those for the numeric data. In the analyses we conducted for numeric and ordinal data, as with nominal data, the choice of the ordering heuristic had more impact on the displays obtained than the choice of distance measures. The ordering solution of some heuristics, such as the hierarchical clustering techniques of single linkage, complete linkage, and the edge method introduced here, uses the rank positioning of the pair-wise distances between the singlet strings during agglomerative clustering without computing composite inter-cluster distances. This makes solutions from these techniques invariant in the set of all monotonic functions of the distance measure used. A property which leads to a marked robustness to the choice of distance measures and the unlikelihood of a marked improvement using alternatives among reasonable distance measures.

In the next section, we look at a tool that resolves the pair-wise mutual distances between data strings into a higher-dimensional space to achieve an ordering of the data.

19.2.2 Multi-dimensional Scaling

To motivate this approach, consider data strings A, B, and C. If we find that the distance measure between A and B is 3, between A and C is 4, and between B and C is 5, then these distances cannot be resolved in one dimension. The numbers will be familiar to anyone who has had a grade school geometry class—A, B, and C can be placed on a plane at the vertices of a right-angled triangle with B and C at the ends of the hypotenuse and A at the right angle. We needed a second dimension to resolve the distances. Multi-dimensional scaling (MDS), is a multivariate tool which uses all pair-wise distances between N data strings and resolves these distances in $N - 1$ dimensions. In the state sequence visualization, the N data strings are subject sequences which need to be placed on rows along the y-axis—essentially a mapping from this $(N - 1)$ dimensional space to the 1-dimensional y-axis. When doing this, as in the right panel of Fig. 19.1, we sacrifice $(N - 2)$ dimensions of information and use the most informative dimension extracted by MDS. In Fig. 19.3, the arrow

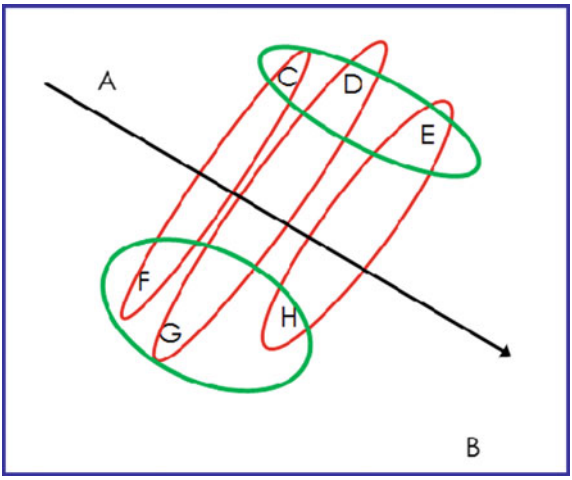


Fig. 19.3 Ordering using a single dimension of a multi-dimensional scaling resolution of pair-wise distances and a contrast with the use of clustering

represents the dimension which appears to capture maximal information as the points seem to be most aligned along that slant. Ordering of the data using MDS would be based on projections on this dimension. If each point represented a subject with a sequence string then subject C would have his color strip representing his state transitions in a row very close to F, D very close to G and E close to H with A and B at the top and bottom rows.

In contrast an ordering based on hierarchical clustering would likely have subjects C, D, and E on adjacent rows of the sequence plot and the F, G, and H subject strips will be contiguous as well—a solution, which in this example, seems to reflect inter subject distances better than the MDS one-dimensional ordering. We will now look at standard hierarchical clustering tools.

19.2.3 Hierarchical Clustering

In this technique, a distance measure between data strings is used along with a clustering rule to form clusters. Table 19.2 illustrates a clustering sequence. In step 1, we have seven data strings in the elements S1 to S7. We start with each single element as a cluster. The two closest elements form a cluster as in step 2 in the table, where elements S1 and S2 from a doublet cluster D1. All distances between elements/clusters are reassessed at each step. At any step, the elements may constitute doublets, triplets, larger clusters or remaining singlets, and the closest distance leads to a new cluster and the process continues till we combine all the data into one cluster. Step 3 and step 4 show some initial steps in the agglomerative clustering of the seven elements in Table 19.2.

Table 19.2 Example of a hierarchical clustering sequence

Step 1 (All Singlets)	Step 2 (1 Doublet, Rest Singlets)	Step 3 (2 Doublets, Rest Singlets)	Step 4 (1 Triplet, 1 Doublets, Rest Singlets)
S1	D1{S1&S3}	D1{S1&S3}	D1{S1&S3}
S2	S2	D2{S2&S7}	T1{S2&S7&S6}
S3			
S4	S4	S4	S4
S5	S5	S5	S5
S6	S6	S6	
S7	S7		

Popular types of hierarchical clustering include single linkage, average linkage, and complete linkage. We hope to add a new hierarchical clustering tool called edge clustering to the lexicon. These clustering techniques start in an identical manner at the initial step when the first doublet cluster is formed. After this step, the techniques diverge as there are different rules for determining distances between clusters when one of them is non-singlet.

Figure 19.4 depicts single linkage clustering. Cluster distance is obtained as the distance between the closest elements across two clusters. In the figure, the next step would be a cluster combining {A, B, D} with {E, G, H} as these clusters are closest by the single linkage inter-cluster distance rule. Figure 19.5 depicts complete linkage clustering, where the inter-cluster distance is computed as the farthest distance between elements of two clusters. Using the complete linkage rule, {A, B, D} would combine with {C, F}. Figure 19.6 illustrates average linkage clustering where the distance between two clusters is defined as the average over all possible pairs of elements chosen from the clusters. Another popular method we considered is the Ward method which minimizes within cluster variances to obtain tight spherical clusters.

These popular hierarchical clustering methods, as conventionally presented, do not provide a unique ordering of clusters or of elements within clusters. The emphasis is on classification. Sakai et al. (2014) obtain a further ordering by sorting, as a separate step, the solution from these classificatory hierarchical clustering methods.

19.2.4 Sorting of Dendrograms from Hierarchical Clustering

Conventional hierarchical clustering methods provide a summary of a classification into clusters using a dendrogram as in Fig. 19.7. This figure contains the IDs of the elements clustered along the x-axis and each staple represents a combination of elements entering from the left and right side of the staple to form a new cluster. One could place the elements, in this ordering along the x-axis, in the rows of a sequence plot, or in the rows or columns of a two-way heat map if the elements represent

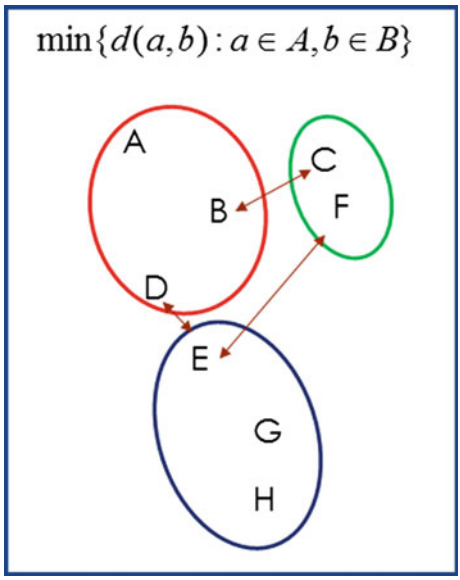


Fig. 19.4 Single linkage clustering rule

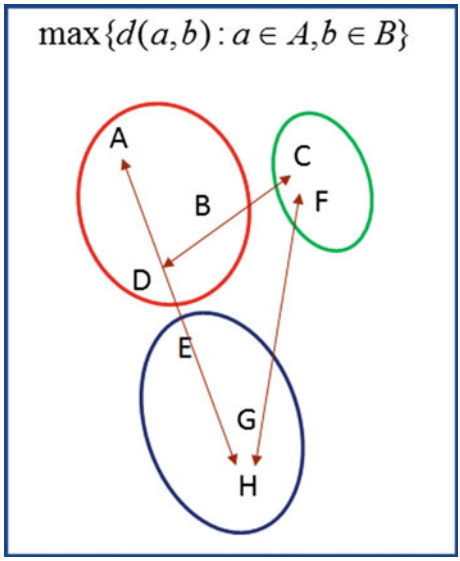


Fig. 19.5 Complete linkage clustering rule

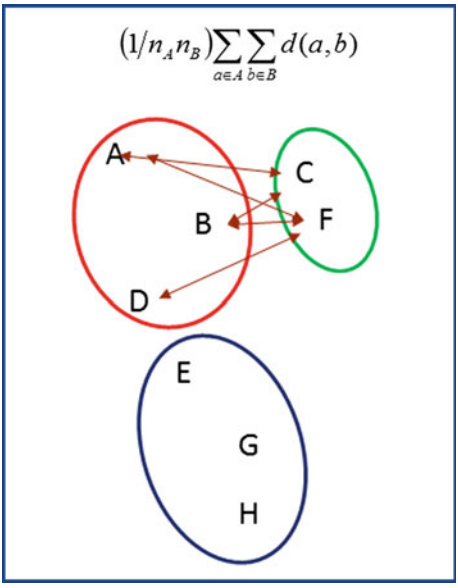


Fig. 19.6 Average linkage clustering rule

row and column strings of data, respectively. However, note that the ordering is not unique. Any combination of the staples in this dendrogram could be reversed to yield a new ordering while preserving the cluster solution.

Sakai et al. (2014) obtain a unique ordering by sorting the dendrogram obtained from a classificatory clustering method as a separate step following the hierarchical clustering. They discuss their method, a 'leaf' ordering following hierarchical clustering, along with other methods involving such a two-step process (Bar-Joseph et al. 2001; Gruvaeus and Wainer 1972). These latter methods are available through the seriation R Package (Buchta et al. 2008). The framework currently used for evaluating estimated heat maps, obtained by using these leaf ordering methods on cluster solutions, uses the information contained in the estimated heat map alone. For instance, Sakai et al. (2014) note the efficiency of their ordering using the data-ink ratio (Tufte 2001), a method which measures the total length of lines required to draw the dendrogram associated with the heat map. Instead of using normative criteria derived solely from estimated heat maps, we will use the dual parameter and estimate statistical framework to compare leaf ordering after clustering versus edge clustering. The edge clustering tool incorporates sorting into the agglomerative clustering heuristic.

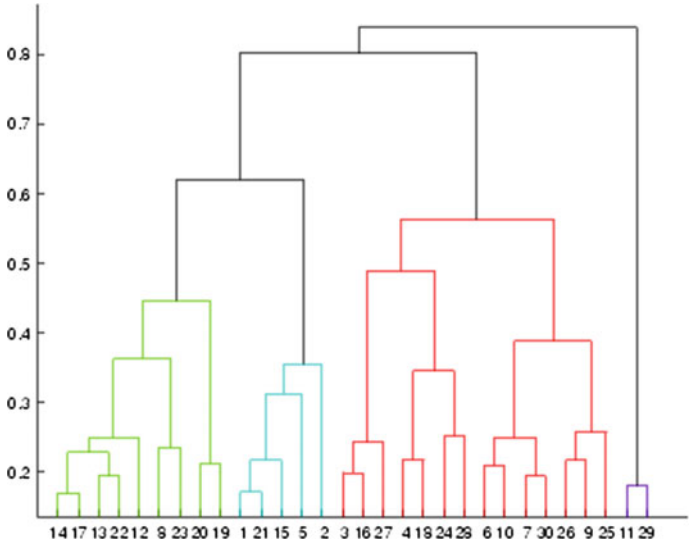


Fig. 19.7 Example of a dendrogram

19.2.5 The New Edge Hierarchical Clustering Method

In this technique, we start with the data strings to be placed in either the rows or columns of the end graphical display. Given the eventual placement of these data strings in rows and columns, and to bring out constraints of such a requirement, we will refer to these data strings as strips. As with all hierarchical clustering methods, the initial step forms the initial doublet cluster strip using the two closest elements among the singlet strips. After this step, the distance measure between two cluster strips is defined as the smallest among the four distances between the data strips at the long edges of the two cluster strips.

In Fig. 19.8, the smallest among the four distances between [A, D] and [C, F] in the ordered clusters {A, B, D} and {C, F} will be compared with all other such inter-cluster distances to determine the clusters to combine in the next clustering step. A new cluster strip is formed by joining the two closest cluster strips at the closest edges. When we continue agglomerating such ordered cluster strips, we end up with a clustering solution as well as an ordering of all data strips into rows or columns. One then needs to map the values in the data strings to elements of a color palette to obtain the visualization.

R Function code to do edge clustering is provided in Appendix A. By ordering on similarity along edges of cluster strips, this method ensures a smoothness in the end graphic. Note that the clustering dendrogram solution will tend to differ from those in other hierarchical methods discussed in Sect. 19.2.3 as the edge method does not use data strips which are not at the edges of a cluster when evaluating inter-cluster

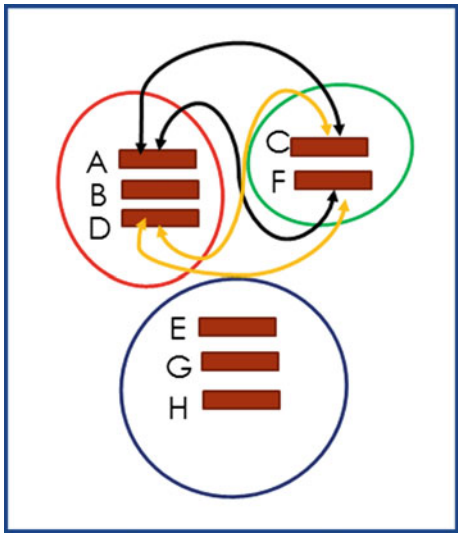


Fig. 19.8 Edge clustering rule

distance. The single, complete, and average linkage methods consider all elements of clusters when evaluating inter-cluster distances. In the next section, we will evaluate the edge clustering heuristic against leaf ordering after other hierarchical clustering methods.

19.3 Assessment of Edge Clustering

We will evaluate the edge clustering heuristic against selected leaf-ordered clustering solutions from other hierarchical methods. The evaluation will involve the use of a statistical parameter-estimate framework as described next.

19.3.1 Assessment Approach: Recovery from Randomly Permuted Known Informative Images

We presume there is an unknown latent image in our data strings which is worth uncovering through our sequence transition plots and heat maps. This is our 'parameter'. Heat maps using different methods are our estimators. The traditional statistical framework to evaluate the efficiency of estimators involves generating data using known parameters and then comparing different parameter blinded estimates derived using just the data, to the parameters. Here we will start with some known informa-

tive image, obtain numeric data as color intensities within pixels, randomly permute rows and columns to remove all ordering information in the image, and attempt to recover the known informative image using the different ordering heuristics.

Such a known image is like row and column data strings, in a two-way heat map, with the numeric data mapped to color intensities. Gene expression data has rows with samples, columns with genes and a numeric gene expression at each row column intersection. Known informative images contain an x pixel coordinate (like genes in columns), a y pixel co-ordinate (like samples in rows) and three numeric values for the intensities of the red, blue, and green colors at that coordinate. This can be converted to monochrome and one intensity value (like numeric gene expression) using the following R code and the imager package:

```
library(imager)
color <- load.image("H:/Stat articles/two way
cluster/First_ladies.jpg")
bw<- grayscale(color)
bw_data <- as.data.frame(bw)
```

The data frame bw_data above has the same structure as data frames for gene expression. We will now look at some iconic images, convert them to numeric 'heat map' like data, permute to lose ordering information and then attempt to recover the images using the ordering heuristics we have discussed. We will start with the photograph of the first ladies meeting at the white house after the 2016 elections.

19.3.2 The First Ladies at the White House

Figure 19.9 has the grayscaled photograph of the first ladies as well as the image after permuting the columns and rows of the grayscaled image. Note that the image has 246 × 166 cells. The R code to permute the rows and columns is provided below. Such a random permutation in gene expression data may move, among other moves, a sample D in row 8 and a gene Y in column 26, in what might be the 'right' heat map, to say row 61 and column 17. The normalized gene expression value corresponding to sample D and gene Y of say 1.73 would now be in cell {61, 17} instead of cell {8, 26} in what is likely a very non-informative heat map.

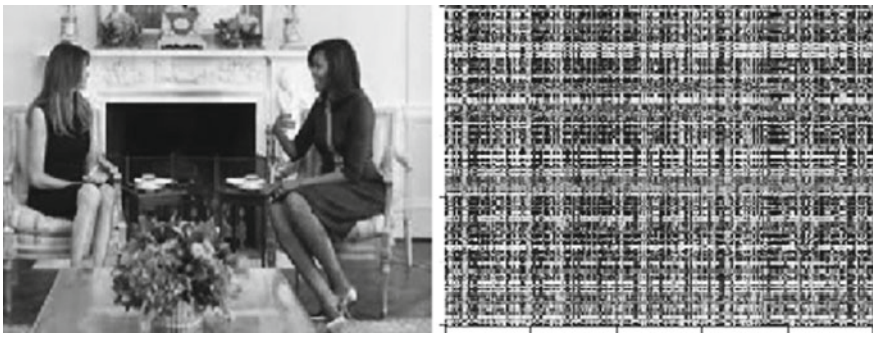


Fig. 19.9 Grayscaled first ladies (left) and permuted image (right)

```
VEC <- bw_data$value
VECt <- matrix(VEC,246,166)
#Randomly permute the row order
set.seed(1234567)
ind <- 1:166
Rind <- sample(ind,length(ind),replace=FALSE,prob=NULL)
#Replace original ordered row numbers with permuted row
numbers
rVECt <-VECt[Rind,]
#Randomly permute the column order
set.seed(145967)
ind <- 1:246
Cind <- sample(ind,length(ind),replace=FALSE,prob=NULL)
#Replace original ordered column numbers with permuted column
numbers
rcVECt <-rVECt[,Cind]
image(t(rcVECt),col=paste("gray",1:99,sep=""))
```

'rcVECt' contains the permuted row and columns data strings for our grayscaled image. Notice that we replaced the correct column and row numbering of this data with an arbitrary string of columns and row numbers.

We will use the information contained in these permuted row and column data strings to estimate an appropriate ordering of the data, replace the permuted ordering with the assessed ordering in the data matrix, map the numeric data in the data matrix with this computed ordering onto a grayscale intensity, and plot. The results using the edge hierarchical clustering method are to the left in Fig. 19.10 and the one using leaf ordering of single linkage hierarchical clustering is to the right.

The edge ordering method recovers the information in the first ladies photograph much more effectively than the leaf ordering method. The R code to achieve the ordering using the edge clustering method is provided below. The edgeOrdering function is provided in Appendix A as noted earlier. As a final step before plotting, we replaced the permuted column and row numbering in the data matrix with the edge method estimated ordering of columns and rows.

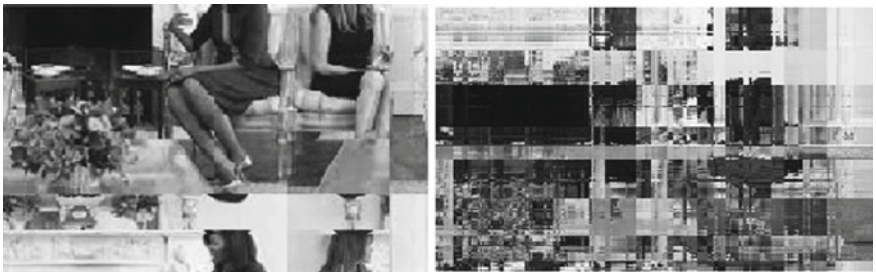


Fig. 19.10 Recovered image using edge clustering (left) and recovered image using leaf ordering of single linkage hierarchical clustering (right)

```
#RowD contains distances between every pair of rows.
#ColD is a similar dataframe for columns.
#edgeOrdering is a function which does the edge ordering.
R_Order <- edgeOrdering(RowD)
C_Order <- edgeOrdering(ColD)
OVECt <-rcVECt[R_Order,C_Order]
#following produces the left panel of Figure 9
image(t(OVECt),col=paste("gray",1:99,sep=""))
```

The R code to conduct the leaf ordering of the single linkage solution is provided below. As noted earlier, this is a two-step process. First, distance measures between the column data strings and distance measures between the row data strings are used to obtain the single linkage clustering dendrograms. Then, these dendrograms are sorted by the dendsort R package to provide the estimated row and column ordering. Before plotting, we replace the permuted column and row numbering in the data matrix with the ordering of columns and rows obtained using the leaf ordering method.

```
#hierachical clustering single linkage
distR <- dist(DFR)
distC <- dist(DFC)
hc0r <- hclust(distR, method="single")
hc0c <- hclust(distC, method="single")
#sort dendrogram
dd0r <- dendsort(as.dendrogram(hc0r))
dd0c <- dendsort(as.dendrogram(hc0c))
hc_sorted0r <- as.hclust(dd0r)
hc_sorted0c <- as.hclust(dd0c)
R_Sin <- hc_sorted0r$order
C_Sin <- hc_sorted0c$order
SVECt <-rcVECt[R_Sin,C_Sin]
#following produces the right panel of Figure 9
image(t(SVECt),col=paste("gray",1:99,sep=""))
```

Fig. 19.11 Grayscaled to left, recovered by the edge method in the middle frame and attempted recovery using leaf ordering after Ward's Hierarchical Clustering in the frame to the right

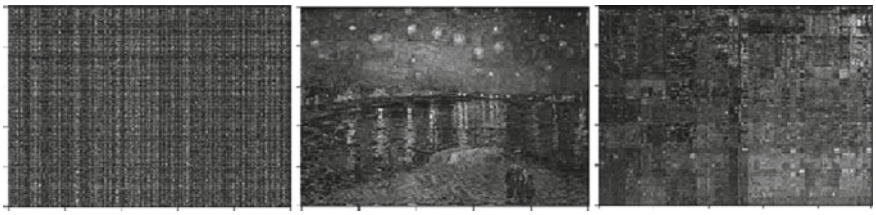


Fig. 19.12 Permuted to left, recovered by the edge method (middle frame) and leaf ordering after average linkage hierarchical clustering (right)

Leaf ordering from complete linkage and average linkage methods was no more effective in recovering the first ladies image than the single linkage method. Additional code and the end graphics are provided at (http://www.resourcetepee.com/). Inspired by the discombobulated first ladies, the next section uses a Picasso painting.

19.3.3 Picasso and Van Gogh

In Fig. 19.11, to the left, we have Picasso's Portrait of Dora Maar (1937). The middle frame has the image recovered after randomly permuting rows and columns using edge clustering. Note that a portion of the wall to the left of the lady in the original moves to the right in the edge recovered image. To the right is a new improved Picasso using leaf ordering on a Ward's cluster solution, which, is on sale by the author at Sotheby's for $10 million!

In Fig. 19.12, we tested edge clustering using Van Gogh's starry night over Rhone.

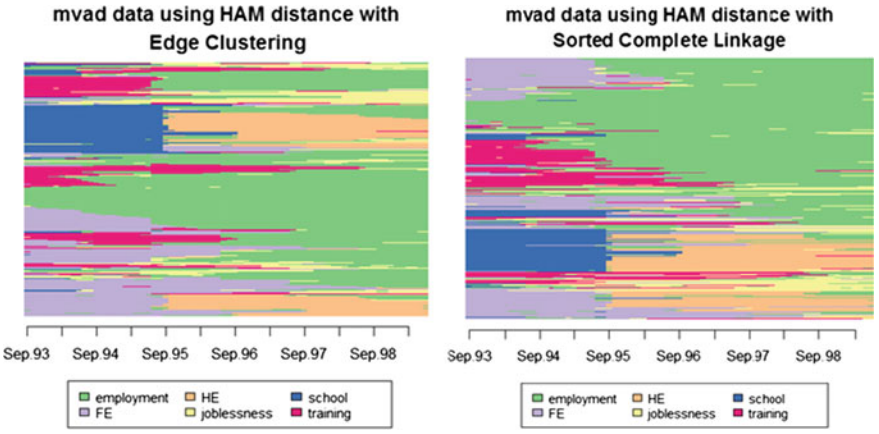


Fig. 19.13 Employment and education sequences ordered by edge clustering (left) and sorted complete linkage (right)

19.4 Examples of the Use of Edge Clustering for Informative Visualization of Real World Data

We now return to some of the applications that we had introduced earlier, starting with the data on transitions between education and employment states (McVicar and Anyadike-Danes 2002).

19.4.1 State Sequence Data

In Fig. 19.1, in the introductory section, we had presented unordered sequence data and data ordered by using multi-dimensional scaling. The multi-dimensional scaling approach was described in Sect. 19.2.2. Distance measures for discrete states such as the HAM measure were discussed in Sect. 19.2.1. The left panel of Fig. 19.13 orders the sequence data using the edge clustering and the right panel uses leaf ordering of the complete linkage cluster solution.

In general, the edge method resulted in graphics which were smoother than those using leaf ordering. However, note the similarity in gross features between the two graphics. This validates the edge clustering tool—it has been developed within hierarchical clustering and tested in a statistical evaluation framework, and is not an atheoretical data crunching or image processing heuristic. The graphics were generated using the following code:

```
#call additional required R packages
library(permute)
library(vegan)
library(dendsort)
#format distance measures for use in the edgeOrdering func-
tion
HAMdist1 <- max(HAMdist)-HAMdist
HAMdistance<-cbind(rep(1:712,each=712),rep(1:712,712),
as.vector(HAMdist1))
HAMdistance1 <- HAMdistance [(HAMdistance [,2]
> HAMdistance[,1]),]
colnames(HAMdistance1) <- c("base", "compare", "sact")
HAMdistance1 <- as.data.frame(HAMdistance1)
# edge clustering and sequence plotting
edgeSeqHAM <- edgeOrdering(HAMdistance1)
mvad_seq_HAME <- mvad.seq[edgeSeqHAM, ]
seqIplot(mvad_seq_HAME,sortv=712:1,cex.legend=0.90,
lab=NA,yaxis=FALSE, title="mvad data using HAM distance with
Edge Clustering")
```

The subject states in this section involved discrete states which were nominal. Next, we look at transitions between discrete states which can be ordered, such as those of response categories, progression and death in cancer patients.

19.4.2 Sequences in Oncological States

The example in this section is based on blinded data from an oncology clinical trial. Transitions between oncological states are compared across subjects in two randomized treatment groups consisting of induction therapy for the period plotted in Fig. 19.14, followed by differing maintenance. Response to therapy during the induction phase was expected to be similar. Oncological states in the longitudinal plot included Complete Response (CR—1), Very Good Partial Response (VGPR—2), Partial Response (PR—3), Stable Disease (SD—4), Progressive Disease (PD—5), and Death (6). The associated numbers reflect the ordinality of the data. After a documented PD, all states were labeled PD till any death, and after death, all states are labeled death. All data on responses better than PD were separated from the PD and death data, and imputations were done using the ordinal logistic regression method in the MICE R package. The imputed dataset was back-merged to the PD and death data with any imputed response states overwritten by the PD and death states. The Euclidean distance measure was used, followed by edge clustering to order the rows appropriately before plotting. As expected, progression-free survival and overall survival curves during induction did not separate as essential differences between randomized groups were at maintenance. The graphics, however, do bring

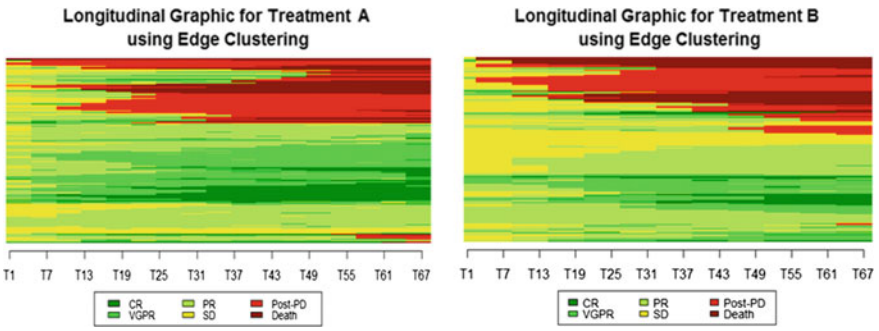


Fig. 19.14 Transitions between oncological states for two induction therapies

out differences between the two induction phases. There was deeper response in the Treatment A panel with more Complete Response and Very Good Partial Response compared to Treatment B where there was more Stable Disease. Many other features, usually summarized separately in multiple data tables, such as time to response, duration of response, time to progressive disease and time to death are brought out in this one display.

The next example looks at displays which require an ordering on both rows and columns and typically involve numeric data.

19.4.3 Gene Sample Heat Maps

In this section, we return to the gene sample data in Sect. 19.1.2. We use the edge clustering and the leaf ordering methods to order both the rows and the columns of the Cancer Genome Atlas (TCGA) data (The Cancer Genome Atlas Research Network 2014) provided with the dendsort R package (Sakai 2015). Figure 19.15 shows the heat map without the dendrograms with the clustering solution for the samples and the genes. The panel to the left is obtained by ordering using edge clustering and the one to the right uses leaf ordering and complete linkage clustering. R code for the leaf ordering was based on that provided at (https://rdrr.io/cran/dendsort/f/vignettes/example_figures.Rmd) by Sakia (2015). As with the state sequence graphics, the similarity in the color terrains obtained by edge clustering to that using leaf ordering validates the edge clustering technique.

19.5 Discussion

We used a statistical parameter and estimate framework to evaluate a novel edge hierarchical clustering method. The 'parameter' here is an informative image. The latent informative image for real data is usually unknown, and this leads to difficulty

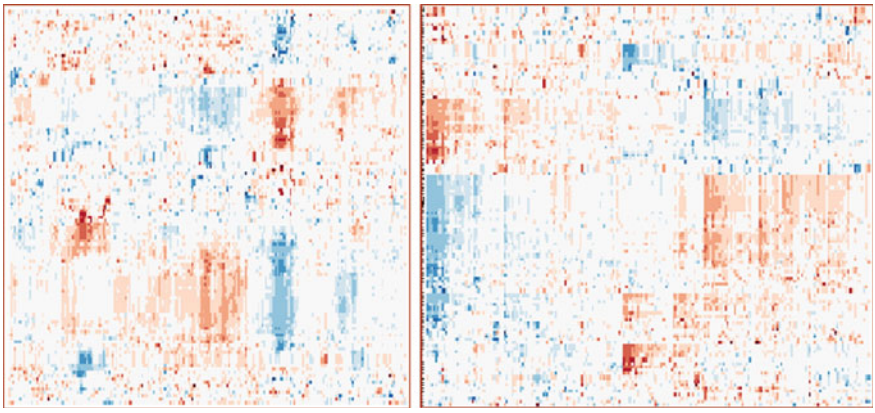


Fig. 19.15 Gene sample heat map using edge clustering (left) and leaf ordering of complete linkage clustering (right)

in ascertaining whether a heuristic is providing a right image. We therefore worked with contexts where images were known to be informative and right. Then, we argued that pixel intensities in such images were like data in heat maps, then permuted rows and columns and attempted to recover the original image. This was achieved with a marked degree of success with edge clustering.

For real data, the leaf ordered hierarchical clustering methods work well, as does the edge clustering heuristic, with some evidence that the edge method may bring out some additional patterns. The graphics exhibited similar features across edge ordering and across leaf orderings of common hierarchical clustering methods. This cross-validates the set of tools studied. One method may bring out a feature in a data set better than another.

We acknowledge that contexts where we somehow know that an image is right and informative (the first ladies photograph, art work) may be different from the ones where we want to find the latent unknown image. By analogy, we may be testing estimators (in the usual context) under, say, a mixture of normal distributions, when our real data will never meet such assumptions. Further, there may be many latent images worth uncovering. Aesthetic images with smoothness qualities, such as those where the edge clustering appears to perform well, may be different from scientifically informative images. However, if you or your translational scientist tries edge clustering and sees Elvis in a heat map, please do let us know!

Acknowledgements The authors would like to thank Arlene Swern and Janice Grecko for their leadership and support of this necessary endeavor to support a tool for the global assessment of response patterns to cancer therapy and for helping in uncovering useful patterns in gene expression data.

Appendix: R*[1] Code to Generate Ordered Sequences of Rows and Columns Using the Edge Clustering Method

```
#########################################################################
#                          Edge Clustering                              #
#                      Version 2, December 2017                         #
#########################################################################

# Main function - edgeOrdering():
# input a dataset containing variables "base", "compare", and "sact".
#     - "base", "compare": row numbers of the original source dataset
#     - "sact": distance/similarity measure between each combination of
#             base row and compare row
# output a vector containing an ordered sequence of row numbers of the
# original source dataset


edgeOrdering <- function(indata){
   indat <- cbind(base=indata$base, compare=indata$compare, sact=indata$sact)
   indat <- indat[order(indat[,3], decreasing = TRUE),]
   n <- dim(indat)[1]
   Nrow <- max(c(indat[,1], indat[,2]))

   # initialization
   left_orig <- left <- right_orig <- right <- list()
   left_orig[1:n] <- left[1:n] <- indat[,1]
   right_orig[1:n] <- right[1:n] <- indat[,2]
   cl_r <- cl_l <- rep(0, n)
   L <- 0

   while(L<Nrow){
      m <- dim(indat)[1]
      if(m==0){break}
      combined <- c(left[[1]], right[[1]])
      L <- length(combined)
      if(m==1){break}

      indatList <- edge(indat, combined, left, right, cl_l, cl_r, m)
      indat <- indatList$rest
      left <- indatList$left
      right <- indatList$right
      cl_l <- indatList$cl_l
      cl_r <- indatList$cl_r
   }
   combined
}
```

[1] *R Core Team 2013.

```
checkList <- function(baseList, compareV){
   lapply(baseList, function(x){
            if(length(x)==length(compareV)){ 1-sum(x != compareV)}
            else{x <- 0}
            })
}

edge <- function(indata, combined, left, right, cl_l, cl_r, m){
   delete <- NULL
   if(cl_r[1]==1){
      delete <- c(delete, (1:m)[checkList(right, right[[1]])==1])
      delete <- c(delete, (1:m)[checkList(left, rev(right[[1]]))==1])

      index1 <- (1:m)[checkList(right, rev(right[[1]]))==1]
      index1 <- index1[index1 != 1]
      right[index1] <- lapply(right[index1], function(x) x <- rev(combined))
      cl_r[index1] <- 1

      index2 <- (1:m)[checkList(left, right[[1]])==1]
      index2 <- index2[index2 != 1]
      left[index2] <- lapply(left[index2], function(x) x <- combined)
      cl_l[index2] <- 1
   }
   else{
      index1 <- (1:m)[checkList(right, right[[1]])==1]
      index1 <- index1[index1 != 1]
      right[index1] <- lapply(right[index1], function(x) x <- rev(combined))
      cl_r[index1] <- 1

      index2 <- (1:m)[checkList(left, right[[1]])==1]
      index2 <- index2[index2 != 1]
      left[index2] <- lapply(left[index2], function(x) x <- combined)
      cl_l[index2] <- 1
   }

   if(cl_l[1]==1){
      delete <- c(delete, (1:m)[checkList(left, left[[1]])==1])
      delete <- c(delete, (1:m)[checkList(right, rev(left[[1]]))==1])

      index3 <- (1:m)[checkList(left, rev(left[[1]]))==1]
      index3 <- index3[index3 != 1]
      left[index3] <- lapply(left[index3], function(x) x <- rev(combined))
      cl_l[index3] <- 1
```

```
    index4 <- (1:m)[checkList(right, left[[1]])==1]
    index4 <- index4[index4 != 1]
    right[index4] <- lapply(right[index4], function(x) x <- combined)
    cl_r[index4] <- 1
  }
  else{
    index3 <- (1:m)[checkList(left, left[[1]])==1]
    index3 <- index3[index3 != 1]
    left[index3] <- lapply(left[index3], function(x) x <- rev(combined))
    cl_l[index3] <- 1

    index4 <- (1:m)[checkList(right, left[[1]])==1]
    index4 <- index4[index4 != 1]
    right[index4] <- lapply(right[index4], function(x) x <- combined)
    cl_r[index4] <- 1
  }

  index5 <- (1:m)[sapply(left, length) == sapply(right, length)]
  for(k in index5){
    if(sum(left[[k]]!=right[[k]])==0) delete <- c(delete,k)
  }

  delete <- unique(c(1, delete))

  if((m-length(delete))==1) rest <- matrix(indata[-delete,], nrow=1)
  else rest <- indata[-delete,]
  left <- left[-delete]
  right <- right[-delete]
  cl_l <- cl_l[-delete]
  cl_r <- cl_r[-delete]

  if((m-length(delete))>1){
    orderNew <- order(rest[,3], decreasing = TRUE)
    rest <- rest[orderNew,]
    left <- left[orderNew]
    right <- right[orderNew]
    cl_l <- cl_l[orderNew]
    cl_r <- cl_r[orderNew]
  }

  list(rest=rest, left=left, right=right, cl_l=cl_l, cl_r=cl_r)
}
```

References

Anderson, T. W. (1958). *Wiley publications in statistics. An introduction to multivariate statistical analysis*. Hoboken, NJ, US: Wiley.

Bar-Joseph, Z., Gifford, D. K., & Jaakkola, T. S. (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics, 17*(Suppl 1), S22–S29.

Buchta, C., Hornik, K., & Hahsler, M. (2008). Getting things in order: An introduction to the R package seriation. *Journal of Statistical Software* 25(3).

Elder, G. H., & Kirkpatrick Johnson, M., & Crosnoe, R. (2003). The emergence and development of life course theory. In *Handbook of the life course* (pp. 3–19). https://doi.org/10.1007/978-0-306-48247-2_1.

Gabadinho, A., et al. (2013, October 11). Workshop on sequence analysis. New York.

Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. S. (2010). Mining sequence data in R with the TraMineR package: A user's guide. University of Geneva. http://mephisto.unige.ch/traminer.

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 40*(4), 1–37. http://www.jstatsoft.org/v40/i04.

Gabadinho, A., Studer, M., Müller, N. S., Bürgin, R., & Ritschard, G. (2016). Trajectory Miner (TraMineR): A Toolbox for Exploring and Rendering Sequences. R package version 1.8-13.

Giele, J. Z., & Elder, G. H., Jr. (Eds.). (1998). *Methods of life course research: Qualitative and quantitative approaches*. Sage Publications. ISBN 0 76191437 4.

Gruvaeus, G., & Wainer, H. (1972). Two additions to hierarchical cluster analysis. *Journal of Mathematical and Statistical Psychology, 25*(2), 200–206.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal, 29,* 147–160.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis*. Englewood Cliffs, N.J: Prentice Hall.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10,* 707–710.

McVicar, D., & Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work using sequence methods. *Journal of the Royal Statistical Society A, 165*(2), 317–334.

Resource Tepee. http://www.resourcetepee.com/.

R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Sakai, R. (2015). dendsort: Modular Leaf Ordering Methods for Dendrogram Nodes. R package version 0.3.3. http://CRAN.R-project.org/package=dendsort.

Sakai, R., Winand, R., & Verbeiren, T. et al. (2014). dendsort: Modular leaf ordering methods for dendrogram representations in R. *F1000Research, 3*, 177.

The Cancer Genome Atlas Research Network. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, Connecticut: Graphics Press.

https://rdrr.io/cran/dendsort/f/vignettes/example_figures.Rmd.

Index

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0