ICSA Book Series in Statistics

Series Editors: Jiahua Chen · Ding-Geng (Din) Chen

Jianchang Lin
Bushi Wang
Xiaowen Hu
Kun Chen
Ray Liu *Editors*

Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics

Selected Papers from the 2015 ICSA/Graybill Applied Statistics Symposium, Colorado State University, Fort Collins

ICSA
泛華統計協會

Springer

ICSA Book Series in Statistics

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

Jianchang Lin • Bushi Wang • Xiaowen Hu
Kun Chen • Ray Liu
Editors

Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics

Selected Papers from the 2015 ICSA/Graybill Applied Statistics Symposium, Colorado State University, Fort Collins

Springer

Editors

Jianchang Lin
Global Statistics
Takeda Pharmaceuticals
Cambridge, MA, USA

Bushi Wang
Biostatistics
Boehringer Ingelheim Pharmaceuticals
Ridgefield, CT, USA

Xiaowen Hu
Colorado State University
Fort Collins, CO, USA

Kun Chen
University of Connecticut
Storrs, CT, USA

Ray Liu
Global Statistics
Takeda Pharmaceuticals
Cambridge, MA, USA

ISSN 2199-0980   ISSN 2199-0999 (electronic)
ICSA Book Series in Statistics
ISBN 978-3-319-42567-2   ISBN 978-3-319-42568-9 (eBook)
DOI 10.1007/978-3-319-42568-9

Library of Congress Control Number: 2016954077

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The 2015 Joint 24th International Chinese Statistical Association (ICSA) Applied Statistics Symposium and 13th Graybill Conference was successfully held from June 14 to June 17, 2015, in Fort Collins, Colorado, USA. The conference covers a variety of exciting and state-of-the-art statistical application topics (over 400 presentations) from the bio-pharmaceutical applications, e.g., clinical trials and personalized medicine, to non-bio-pharmaceutical applications, e.g., finance and business analytics, with attendees from industry, government, and academia.

The 24 papers were selected from the presentations in the annual meeting with a broad range of topics so that readers of this book could not only enjoy the topics close to their own research areas but also from other different areas. All papers have gone through the peer review process and parts covered in the book include:

- biomarker and personalized medicine
- Bayesian methods and applications
- dose ranging studies in clinical trials
- innovative clinical trial designs and analysis
- clinical and safety monitoring in clinical trials
- statistical applications in nonclinical and preclinical drug development
- statistical learning methods and applications with large-scale data
- statistical applications in business and finance

We are very grateful to the authors who contributed their papers to these proceedings and carefully prepared their manuscripts within a tight timeline. These proceedings would also not be possible without the successful symposium, which gave us the opportunity to share, learn, and choose so many high-quality papers. Our deep gratitude goes to the leadership of Naitee Ting and the executive organizing committee, the program committee, and many other volunteers of the

24th ICSA Applied Statistics Symposium and 13th Graybill Conference. We also thank Michael Penn of Springer for the assistance through the entire process of completing the book.

Cambridge, MA, USA	Jianchang Lin
Ridgefield, CT, USA	Bushi Wang
Fort Collins, CO, USA	Xiaowen Hu
Storrs, CT, USA	Kun Chen
Cambridge, MA, USA	Ray Liu

Contents

Contributors

Robert Aseltine Division of Behavioral Science and Community Health, University of Connecticut Health Center, Farmington, CT, USA

Center for Public Health and Health Policy, University of Connecticut Health Center, Farmington, CT, USA

Greg Ball Biostatistics and Research Decision Sciences, Merck Research Laboratories, Rahway, NJ, USA

Bryan Bernat Hospira, a Pfizer company, Lake Forest, IL, USA

Mário de Castro Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brazil

Hao Chai Department of Biostatistics, Yale University, New Haven, CT, USA

Ming-Hui Chen Department of Statistics, University of Connecticut, Storrs, CT, USA

Kun Chen Department of Statistics, University of Connecticut, Storrs, CT, USA

Center for Public Health and Health Policy, University of Connecticut Health Center, Farmington, CT, USA

Chi-Tian Chen Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan

Alan Y. Chiang Global Statistical Sciences, Lilly Corporate Center, Eli Lilly and Company, Indianapolis, IN, USA

Sy Han Chiou Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Katherine Curtis Department of Community and Environmental Sociology, University of Wisconsin-Madison, Madison, WI, USA

Qiqi Deng Biostatistics and Data Sciences, Boehringer-Ingelheim Pharmaceuticals, Inc., Ridgefield, CT, USA

Dipak Dey Department of Statistics, University of Connecticut, Storrs, CT, USA

Center for Public Health and Health Policy, University of Connecticut Health Center, Farmington, CT, USA

Hong-Bin Fang Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University Medical Center, Washington, DC, USA

Joseph C. Gardiner Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA

Gregory J. Hather Takeda Pharmaceuticals, Cambridge, MA, USA

Gordon Honerkamp-Smith Department of Mathematics, University of California, San Diego, CA, USA

Chin-Fu Hsiao Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan

Hengzhen Huang Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University Medical Center, Washington, DC, USA

Shu-Fu Kuo Division of Biostatistics and Bioinformatics, National Health Research Institutes, Zhunan Township, Miaoli County, Taiwan

Department of Mathematics, National Chung Cheng University, Minxiong, Chiayi County, Taiwan

K.K. Gordon Lan Janssen Pharmaceutical Companies of Johnson & Johnson, Raritan, NJ, USA

Jianjun (David) Li Pfizer, Inc., Collegeville, PA, USA

Jianchang Lin Takeda Pharmaceuticals, Cambridge, MA, USA

Li-An Lin Merck Research Laboratories, Rahway, NJ, USA

Pei-Sheng Lin Division of Biostatistics and Bioinformatics, National Health Research Institutes, Zhunan Township, Miaoli County, Taiwan

Department of Mathematics, National Chung Cheng University, Minxiong, Chiayi County, Taiwan

Stuart Lipsitz Division of General Medicine, Brigham and Womens Hospital, Boston, MA, USA

Jung-Tzu Liu Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan

Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu, Taiwan

Yuefeng Lu Biostatistics and Programming, Sanofi, Framingham, MA, USA

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

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

VA Palo Alto Health Care System, Palo Alto, CA, USA

Chongliang Luo Department of Statistics, University of Connecticut, Storrs, CT, USA

Center for Public Health and Health Policy, University of Connecticut Health Center, Farmington, CT, USA

Zhehui Luo Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA

Shuangge Ma Department of Biostatistics, Yale University, New Haven, CT, USA

Xiwen Ma Biostatistics and Programming, Sanofi, Framingham, MA, USA

Brian A. Millen Eli Lilly and Company, Lily Corporation Center, Indianapolis, IN, USA

Richard Montes Hospira, a Pfizer company, Lake Forest, IL, USA

Adriano Polpo Department of Statistics, Federal University of São Carlos, São Carlos, SP, Brazil

Serap Sankoh Takeda Pharmaceuticals, Cambridge, MA, USA

Patrick M. Schnell Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA

Leslie Sidor Biogen, Cambridge, MA, USA

Debajyoti Sinha Department of Statistics, Florida State University, Tallahassee, FL, USA

Yanyan Song Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA

Department of Pharmacology and Biostatistics, Institute of Medical Sciences, Shanghai Jiao Tong University, Shanghai, China

Catherine Srebalus-Barnes Hospira, a Pfizer company, Lake Forest, IL, USA

Ming T. Tan Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University Medical Center, Washington, DC, USA

Xiaoqin Tang Asthma, Allergy and Autoimmunity Institute, Allegheny Health Network, Pittsburgh, PA, USA

Zhaoyang Teng Takeda Pharmaceuticals, Cambridge, MA, USA

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

Department of Statistics, Stanford University, Stanford, CA, USA

Naitee Ting Biostatistics and Data Sciences, Boehringer-Ingelheim Pharmaceuticals, Inc., Ridgefield, CT, USA

Hsiao-Hui Tsou Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan

Graduate Institute of Biostatistics, College of Public Health, China Medical University, Taichung, Taiwan

Chyng-Shyan Tzeng Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu, Taiwan

Gregory Vaughan Department of Statistics, University of Connecticut, Storrs, CT, USA

Ming-Dauh Wang Global Statistical Sciences, Lilly Corporate Center, Eli Lilly and Company, Indianapolis, IN, USA

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

Ronghui Xu Department of Mathematics, University of California, San Diego, CA, USA

Department of Family Medicine and Public Health, University of California, San Diego, CA, USA

Jin Xu Merck Sharp & Dohme, Kenilworth, New Jersey, PA, USA

Jun Yan Department of Statistics, University of Connecticut, Storrs, CT, USA

Center for Public Health and Health Policy, University of Connecticut Health Center, Farmington, CT, USA

Yang Yang Division of Biometrics 1, Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

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

Ying Yuan Department of Biostatistics, MD Anderson Cancer Center, Houston, TX, USA

Yong Zang Department of Biostatistics, Indiana University, Indianapolis, IN, USA

Yangguang Zang School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China

Zheng Zhang Department of Biostatistics, Center for Statistical Sciences, Brown University School of Public Health, Providence, RI, USA

Bin Zhang Genocea Biosciences, Cambridge, MA, USA

Qingzhao Zhang Department of Biostatistics, Yale University, New Haven, CT, USA

Yuanye Zhang Agios Pharmaceuticals, Cambridge, MA, USA

Sanguo Zhang School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China

Yinjun Zhao Department of Biostatistics, Yale University, New Haven, CT, USA

Wei Zheng Biostatistics and Programming, Sanofi, Framingham, MA, USA

Jun Zhu Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA

Part I
Biomarker and Personalized Medicine

Optimal Biomarker-Guided Design for Targeted Therapy with Imperfectly Measured Biomarkers

Yong Zang and Ying Yuan

Abstract Targeted therapy revolutionizes the way physicians treat cancer and other diseases, enabling them to adaptively select individualized treatment according to the patient's biomarker profile. The implementation of targeted therapy requires that the biomarkers are accurately measured, which may not always be feasible in practice. In this article, we propose two optimal biomarker-guided trial designs in which the biomarkers are subject to measurement errors. The first design focuses on a patient's individual benefit and minimizes the treatment assignment error so that each patient has the highest probability of being assigned to the treatment that matches his/her true biomarker status. The second design focuses on the group benefit, which maximizes the overall response rate for all the patients enrolled in the trial. We develop a likelihood ratio test to evaluate the subgroup treatment effects at the end of the trial. Simulation studies show that the proposed optimal designs achieve our design goal and obtain desirable operating characteristics.

Keywords Biomarker-guided design • Measurement error • Optimal design • Personalized medicine

1 Introduction

With accumulating knowledge on cancer genomics and rapid developments in biotechnology, targeted therapy (or personalized medicine) provides an unprecedented opportunity to battle cancer. Targeted therapy is a type of treatment that blocks the growth of cancer cells by identifying and attacking specific functional units needed for carcinogenesis and tumor growth while sparing normal tissue (Sledge 2005). Targeted therapy is based on the notion that the genetic mechanism of

Y. Zang
Department of Biostatistics, Indiana University, Indianapolis, IN, USA
e-mail: zangyong2008@gmail.com

Y. Yuan (✉)
Department of Biostatistics, MD Anderson Cancer Center, Houston, TX, USA
e-mail: yyuan@mdanderson.org

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_1

cancer is heterogeneous across patients. In order to treat patients more effectively, the treatment should be matched to the individual's genetic profile or biomarker status (e.g., a certain gene mutation or oncologic pathway activation).

The biomarker-guided design (Mandrekar and Sargent 2009; Freidlin et al. 2010) provides an essential framework for determining whether the agents under investigation are effective in the corresponding marker subgroups, compared to the effectiveness of untargeted treatments in historical controls. Under this design, when a new patient is enrolled, we first measure his/her biomarker, based on which we then adaptively assign the patient to one of the targeted treatments that matches the patient's marker status.

An essential requirement for using the biomarker-guided design is that, after a patient is enrolled, we are able to quickly and accurately assess his/her marker status and then use that information to assign him/her to an appropriate treatment in a timely fashion. Modern high-throughput methods, such as microarrays and next-generation sequencing technology, provide accurate and high-fidelity ways to measure a patient's gene profile and biomarker status. However, these methods are time-consuming and logistically complicated. In addition, high-throughput methods are relatively expensive and therefore it may be financially infeasible to apply them to all patients in a trial. To avoid these issues, we can measure patient biomarker status using surrogate marker information, such as immunohistochemistry or histology. These methods are fast and cheap, but are often less reliable and prone to measurement errors, leading to inefficient trial design and biased estimates. One solution to this dilemma is to use a two-stage approach: at stage I, we enroll n_1 patients and measure their biomarkers using both the expensive error-free method and cheap error-prone method; and then at stage II, we enroll additional n_2 patients and measure their biomarkers only using the error-prone method. By doing so, we (partially) avoid the cost and logistic issues associated with measuring all patients using the expensive error-free method. At the same time, we can use the data from the stage I patients to learn the relationship between the error-free measure and the error-prone measure, based on which make appropriate adjustment to assign the stage II patients and obtain consistent estimates. This is the strategy we adopt here.

In this article, we propose two optimal biomarker-guided designs for the scenario in which some patients' biomarkers are measured with the surrogate marker information. The first design focuses on the patients' individual benefit and minimizes the treatment assignment error, so that each patient has the largest probability of being assigned to the treatment that matches his/her true biomarker status. The second design focuses on the group benefit and maximizes the total number of responses in the trial. We propose a likelihood ratio test for subgroup analysis at the end of the trial.

2 Methods

2.1 Optimal Allocation Rules

Let X denote a continuous error-free measure for the marker of interest, which follows a normal distribution $N(\mu_x, \sigma_x^2)$. Based on the value of X, we classify patients into two subgroups: a marker-positive subgroup (denoted by $M = 1$) if $X \geq \tau$ and a marker-negative subgroup (denoted by $M = 0$) otherwise, where τ is a prespecified cutoff (e.g., the median of X). Let $T = 1$ denote the treatment targeting the marker-positive subgroup (i.e., $M = 1$), and $T = 0$ denote the treatment targeting the marker-negative subgroup (i.e., $M = 0$). Let Y denote the binary response outcome, with $Y = 1$ indicating a response. Under the biomarker-guided design, patients are treated according to their marker status. Specifically, for a newly enrolled patient, we first measure his/her marker status M and then assign the patient to treatment $T = 1$ if $M = 1$ and to treatment $T = 0$ if $M = 0$.

Suppose that cost and logistic issues limit the measurement of X to only the first n_1 out of a total of n patients, while an easy-to-obtain but error-prone surrogate marker measure, W, is available for all n patients. We assume that W follows the classical measurement error model (Fuller 1987; Carroll et al. 2006) as $W = \alpha + \beta X + U$, where U is a random error that is independent of X and follows $N(0, \sigma_u^2)$. For convenience, according to how the marker is measured, we divide the trial into two stages: stage I consists of the first n_1 patients, for which both X and W are measured, and stage II consists of the remaining $n_2 = n - n_1$ patients, for which only W is measured.

As X (thus M) is not observed for stage II patients, the difficulty of conducting the biomarker-guided design is determining how to assign these patients to appropriate treatments in real time based on W. To address this issue, we propose an optimal design, denoted as MinError design, that minimizes the probability of incorrect treatment assignment ($\text{pr}(T \neq M|W)$) during the trial conduct. The basis of the MinError design is the following optimal treatment assignment rule.

Theorem 1. *The probability of treatment misassignment* $\text{pr}(T \neq M|W)$ *is minimized by assigning a patient with an error-prone measure W to treatment $T = 1$ if $\pi(W) \leq 1/2$ and otherwise to $T = 0$, where $\pi(W) = \text{pr}(M = 0|W)$ is the predictive probability that the patient's true marker status is negative given the error-prone measure W.*

With this result at hand, we develop the two-stage MinError design. At stage I, we enroll n_1 patients and measure their biomarkers, including the error-free measure X and error-prone measure W. If $X \geq \tau$ (i.e., $M = 1$), we assign the patient to $T = 1$ and otherwise to $T = 0$. At stage II, we enroll additional n_2 patients, and obtain their biomarker measures W. If $\pi(W) \leq 1/2$, we assign the patient to treatment $T = 1$ and otherwise to $T = 0$. In addition, Implementing the MinError design requires the evaluation of $\pi(W)$, which can be done by transforming the classical error model into a regression calibration model as $X = \alpha^* + \beta^* W + U^*$ where U^*

follows a normal distribution $N(0, \sigma_{u*}^2)$. To estimate $\pi(W)$, we can fit stage I data to the regression calibration model to obtain $\hat{\alpha}^*$, $\hat{\beta}^*$, $\hat{\sigma}_{u*}$, and then estimate $\pi(W)$ by $\Phi(\frac{\tau-\hat{\alpha}^*-\hat{\beta}^* W}{\hat{\sigma}_{u*}})$.

The MinError design minimizes the probability that a patient will be misassigned to an incorrect treatment. It can be viewed as a procedure that optimizes the patients' individual benefit. From another perspective, we may regard the patients enrolled in the trial as a group for which we are interested in optimizing the overall benefit, for example, maximizing the overall response rate, i.e., $\text{pr}(Y = 1)$. Let $p_{jk} = \text{pr}(Y = 1|M = j, T = k)$ denote the response rate for patients with marker $M = j$ under treatment $T = k$. Hence, $p_{11} - p_{10}$ (or $p_{00} - p_{01}$) presents the penalty of incorrectly assigning patients with $M = 1$ (or $M = 0$) to treatment $T = 0$ (or $T = 1$). When the targeted therapy are effective in the marker subgroup $M = 1$ (or $M = 0$), we have $p_{11} > p_{10}$ (or $p_{00} > p_{01}$). The following theorem provides the treatment assignment rule that maximizes the overall response rate. We refer to the resulting design as the MaxResp design.

Theorem 2. *Define $\omega = p_{00} + p_{11} - p_{10} - p_{01}$ and $\delta = (p_{11} - p_{10})/(p_{00} - p_{01})$. The overall treatment response rate $\text{pr}(Y = 1)$ is maximized by assigning a patient with an error-prone measure W to treatment $T = \text{I}(\omega > 0)$ if $\pi(W) \leq \delta/(1 + \delta)$, and otherwise to $T = 1 - \text{I}(\omega > 0)$, where $\text{I}(\cdot)$ is the indicator function.*

The MaxResp design has the same structure as the MinError design, except that in stage II, the MaxResp design uses the treatment assignment rule as described in Theorem 2 to assign patients, while the MinError design uses the treatment assignment rule as described in Theorem 1. However, in the case for which $\delta = 1$, the MaxResp design is identical to the MinError design.

2.2 Likelihood Ratio Test Based on EM Algorithm

We have proposed two optimal rules for assigning patients to appropriate treatments during the trial. At the end of the trial, the goal is to determine whether the targeted treatments are effective in the corresponding subgroups. Specifically, assuming a two-sided test, we are interested in the following two subgroup analyses: testing $H_0 : p_{11} = \psi_1$ versus $H_1 : p_{11} \neq \psi_1$ for the $M = 1$ subgroup, and testing $H_0 : p_{00} = \psi_0$ versus $H_1 : p_{00} \neq \psi_0$ for the $M = 0$ subgroup, where ψ_1 and ψ_0 are prespecified response rates. Hereafter, we focus on the treatment arm $T = 1$, noting that the test for the treatment arm $T = 0$ can be done similarly.

We propose a likelihood ratio test based on the EM algorithm (Dempster et al. 1977; Ibrahim 1990) to evaluate the subgroup treatment effect. Let n_{jkl} denote the number of patients allocated to stage j in treatment k with response l ($j = 1, 2; k = 0, 1; l = 0, 1$), and define $n_{jk\cdot} = n_{jk0} + n_{jk1}$. Let y_i, x_i, w_i and m_i denote the response, error-free and error-prone measures and true marker status for the ith patients. We employ the EM algorithm to solve the MLEs of p_{11}, p_{01}, α^*, β^* and σ_{u*}. At the

E-step, we substitute the missing values of m_i with its conditional expectation

$$\frac{p_{11}^{y_i}(1-p_{11})^{1-y_i}(1-\pi(w_i))}{p_{11}^{y_i}(1-p_{11})^{1-y_i}(1-\pi(w_i))+p_{01}^{y_i}(1-p_{01})^{1-y_i}\pi(w_i)}.$$

At the M-step, we update

$$\hat{p}_{11}=\frac{\sum_{i=1}^{n_{11\cdot}} y_i+\sum_{i=n_{11\cdot}+1}^{n_{11\cdot}+n_{21\cdot}} \mathrm{I}(m_i=1)y_i}{n_{11\cdot}+\sum_{i=n_{11\cdot}+1}^{n_{11\cdot}+n_{21\cdot}} \mathrm{I}(m_i=1)},$$

$$\hat{p}_{01}=\frac{\sum_{i=n_{11\cdot}+1}^{n_{11\cdot}+n_{21\cdot}} \mathrm{I}(m_i=0)y_i}{\sum_{i=n_{11\cdot}+1}^{n_{11\cdot}+n_{21\cdot}} \mathrm{I}(m_i=0)},$$

and update $\hat{\alpha}^*$, $\hat{\beta}^*$ and $\hat{\sigma}_{u^*}$ by maximizing $\sum_{i=1}^{n_{11\cdot}+n_{21\cdot}}\{m_i\log(1-\pi(w_i))+(1-m_i)\log(\pi(w_i))\}$. Similarly, the MLEs of p_{01}, α^*, β^* and σ_{u^*} under the null hypothesis can be obtained using the EM algorithm with the constraint $p_{11}=\psi_1$. We denote the resulting MLEs as $\tilde{p}_{01}$, $\tilde{\alpha}^*$, $\tilde{\beta}^*$ and $\tilde{\sigma}_{u^*}$.

With the MLEs at hand, we can build the likelihood ratio test. The observed log-likelihood function can be written as

$$L=\sum_{i=1}^{n_{11\cdot}} y_i\log(p_{11})+(n_{11\cdot}-\sum_{i=1}^{n_{11\cdot}} y_i)\log(1-p_{11})$$
$$+\sum_{i=n_{11\cdot}+1}^{n_{11\cdot}+n_{21\cdot}}\left\{y_j\log(q(w_i))+(1-y_j)\log(1-q(w_i))\right\},$$

where $q(w_i)\equiv \mathrm{pr}(y_i=1|w_i,T=1)=p_{11}\{1-\pi(w_i)\}+p_{01}\pi(w_i)$. Thus, the likelihood ratio test is given by

$$Z=2\{L(\hat{p}_{11},\hat{p}_{01},\hat{\alpha}^*,\hat{\beta}^*,\hat{\sigma}_{u^*})-L(\psi_1,\tilde{p}_{01},\tilde{\alpha}^*,\tilde{\beta}^*,\tilde{\sigma}_{u^*})\}.$$

Given a significance level of ϵ, we reject H_0 if $Z>\chi^2_\epsilon(df=1)$ where $\chi^2_\epsilon(df=1)$ is the upper ϵ quantile of a χ^2 distribution with one degree of freedom.

3 Simulation Studies

We carried out simulation studies to investigate the operating characteristics of the proposed optimal designs. We compared the proposed MinError and MaxResp designs to a naive design that ignores the measurement error for the treatment assignment during the trial conduct (i.e., directly uses W to classify the stage II patients into marker-positive and -negative patients.). In all three designs, the proposed likelihood ratio test was used to test the subgroup treatment effects at the end of the trial.

We simulated the error-free biomarker measure X from $N(0, \sigma_x^2)$ and the error-prone measure W based on the classical measurement error model. We set the threshold $\tau = 0$ so that half of the patients are positive for the marker and half are negative. We assumed that $n_1 = 50$ and $n_2 = 150$ patients were enrolled in stages I and II, respectively. To assess the type I error rate and power of the designs, we considered the null hypothesis that targeted therapies were not effective with $p_{00} = p_{11} = p_{10} = p_{01} = \pi = 0.2$, and the alternative hypothesis (i.e., the targeted therapies were effective) with $(p_{00}, p_{01}, p_{10}, p_{11}) = (0.3, 0.2, 0.2, 0.4)$. We fixed $\alpha = -1$ and investigated different configurations of the measurement error model parameters β, σ_u and σ_x. Under each of the simulation configurations, we conducted 10,000 simulated trials.

Table 1 shows the simulation results, including the treatment misassignment rate, overall response rate, type I error rate and power. Across various simulation settings, the naive design led to the highest misassignment rate among the three designs. As a result, the response rate under the naive design was lower than those under the MinError and MaxResp designs. For example, when $\sigma_x = 0.5$, $\sigma_u = 0.25$ and $\beta = 0.6$, the misassignment rate under the naive design was double those under the MinError and MaxResp designs, and the response rate was about 5 % lower than those of the optimal designs. The empirical type I error rates of the MinError, MaxResp and naive designs were generally close to the nominal level of 5 %, suggesting that the proposed likelihood ratio test effectively accounted for the measurement errors. The MinError and MaxResp designs had the same type I error rates because the designs were equivalent under the null hypothesis.

Compared to the naive design, the MinError and MaxResp designs generally had higher average statistical power to detect the treatment effect. For example, when $\sigma_x = 0.5$, $\sigma_u = 0.25$ and $\beta = 1.0$, the average power of the MinError and MaxResp designs was about 14 % higher than that of the naive design. In general, the two proposed designs, MinError and MaxResp designs, performed rather similarly, although the MinError design had a slightly lower misassignment rate while the MaxResp design had a slightly higher response rate because they optimize different objective functions. Figures 1 and 2 show how the misassignment rate and overall response rate change with the simulation parameters for the three designs. We consistently observed that the MinError design had the lowest misassignment rate and the MaxResp design had the highest response rate.

In Table 1, we fixed $n_1 = 50$, which represents the number of patients whose biomarker profiles are precisely measured. The proposed optimal designs and the naive design perform better when n_1 increased. However, larger n_1 requires an increment of the budget for the biomarker-guided design. Also, in addition to all the proposed design in Table 1, an error-free design which uses the fist stage patients only can be adopted as well. However, this error-free design discards all the information from the patients with imperfectly measured biomarkers. Hence, it is consistently less powerful than the optimal designs.

Table 1 Simulation results for the Naive, MinError and MaxResp designs, with $\alpha = -1$, $n_1 = 50$ and $n_2 = 150$

				Misassignment	Response	Type I error		Power		
σ_x	σ_u	β	Design	rate	rate	$T = 0$	$T = 1$	$T = 0$	$T = 1$	Average
0.5	0.25	0.6	Naive	37.1	27.6	4.7	5.2	48.7	63.0	55.9
			MinError	16.8	32.5	5.4	4.7	36.2	86.3	61.3
			MaxResp	18.6	32.7	5.4	4.7	32.0	90.1	61.1
		0.8	Naive	36.2	27.8	4.8	5.2	53.0	65.9	59.5
			MinError	13.5	33.0	4.9	4.6	41.2	91.2	66.2
			MaxResp	14.9	33.2	4.9	4.6	37.6	93.6	65.6
		1.0	Naive	34.8	28.0	5.3	5.0	55.2	69.8	62.5
			MinError	11.2	33.3	4.9	5.4	55.5	97.9	76.7
			MaxResp	12.3	33.5	4.9	5.4	54.0	98.4	76.2
	0.5	0.6	Naive	35.0	28.0	5.0	5.4	34.8	65.0	49.9
			MinError	25.1	31.3	5.3	5.0	27.0	72.9	50.0
			MaxResp	28.3	31.8	5.3	5.0	22.5	78.3	50.4
		0.8	Naive	33.6	28.3	5.4	5.1	39.9	68.9	54.4
			MinError	21.8	31.8	5.5	4.6	30.0	78.5	54.3
			MaxResp	24.3	32.1	5.5	4.6	24.8	83.8	54.3
		1.0	Naive	32.1	28.6	5.4	4.8	43.6	72.8	58.2
			MinError	19.1	32.2	4.6	4.6	49.3	96.1	72.7
			MaxResp	21.2	32.5	4.6	4.6	45.1	96.9	71.0
1.0	0.25	0.6	Naive	32.9	28.4	5.1	4.7	56.7	74.9	65.8
			MinError	9.5	33.6	5.1	4.6	47.4	95.3	71.4
			MaxResp	10.5	33.7	5.1	4.6	45.2	96.4	70.8
		0.8	Naive	28.8	29.2	4.9	5.1	58.7	82.3	70.5
			MinError	7.3	33.9	5.1	5.4	51.0	96.7	73.9
			MaxResp	8.0	34.0	5.1	5.4	50.1	97.4	73.8
		1.0	Naive	25.0	30.0	5.2	4.5	59.6	88.0	73.8
			MinError	5.9	34.1	5.2	5.5	59.6	98.7	79.2
			MaxResp	6.5	34.3	5.2	5.5	59.4	98.9	79.2
	0.5	0.6	Naive	30.4	28.9	4.8	5.2	46.2	77.3	61.8
			MinError	16.8	32.3	5.4	5.0	36.2	86.3	61.3
			MaxResp	18.6	32.7	5.4	5.0	32.0	90.1	61.1
		0.8	Naive	27.0	29.6	5.3	4.9	50.9	84.1	67.5
			MinError	13.5	33.0	5.5	4.8	41.2	91.2	66.2
			MaxResp	14.9	33.2	5.5	4.8	37.6	93.6	65.6
		1.0	Naive	23.8	30.2	5.1	4.8	54.5	88.5	71.5
			MinError	11.2	33.3	5.3	4.9	55.5	97.9	76.7
			MaxResp	12.3	33.5	5.3	4.9	54.0	98.4	76.2

All values are in percentages

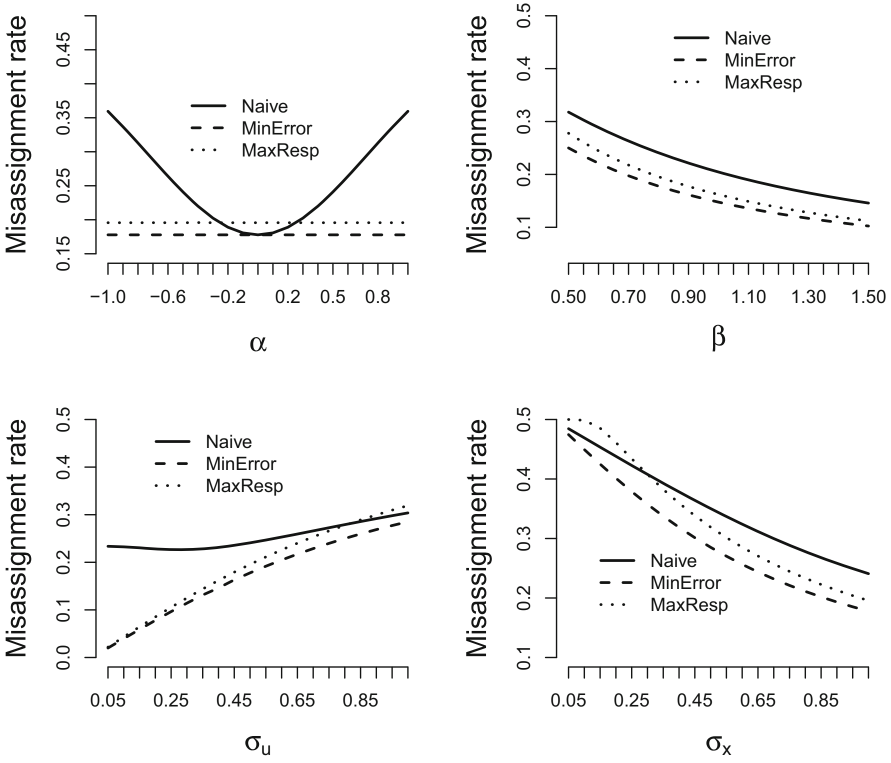


Fig. 1 The misassignment rates of the naive, MinError and MaxResp designs

4 Conclusion

We have proposed two optimal biomarker-guided designs when the biomarkers are subject to measurement errors. The first design focuses on the patients' individual benefit and minimizes the treatment assignment error, so that each patient has the highest probability of being assigned to the treatment that matches his/her true biomarker status. The second design focuses on the group benefit, which maximizes the total number of responses in the trial. We developed a likelihood ratio test to evaluate the treatment effects for marker subgroups at the end of the trial. Simulation studies showed that the proposed optimal designs have desirable operating characteristics. We investigate the binary outcome in this article. It is also of interest to extend the optimal designs by handling other outcomes (e.g., progression-free survival or overall survival). Future research in this area is required.

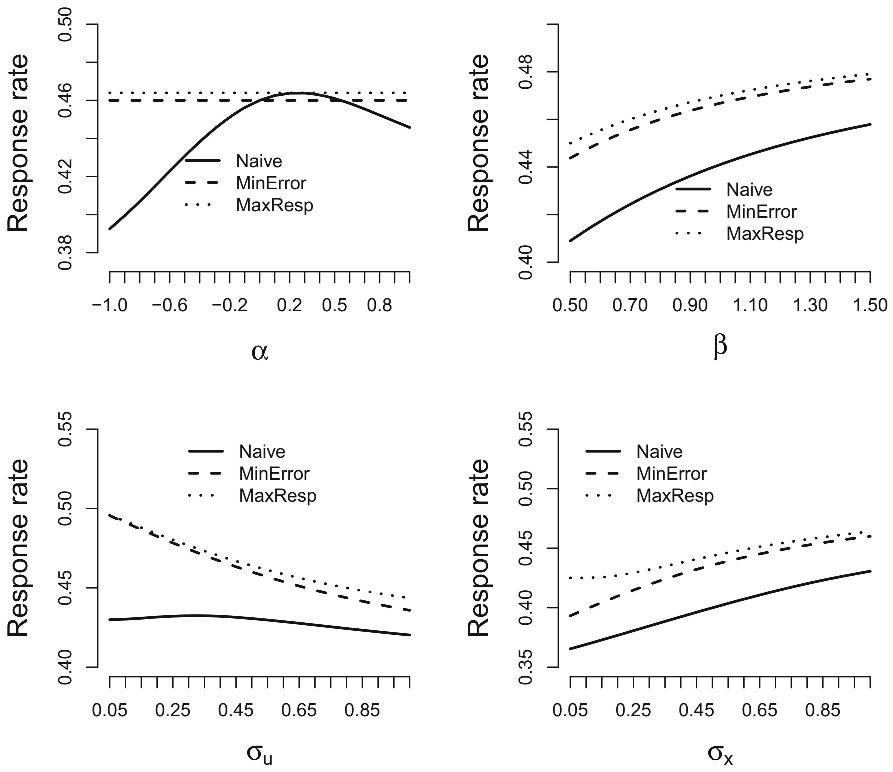


Fig. 2 The response rates of the naive, MinError and MaxResp designs when $p_{10} = p_{01} = 0.1$ and $p_{00} = 0.2$

Appendix

Proof of Theorem 1

For stage II patients, the treatment assignment is solely determined by W, therefore, conditional on W, T and M are independent. It follows that the probability of misassignment for subjects assessed with the error-prone measure W is given by

$$\begin{aligned}
\mathrm{pr}(T \neq M|W) &= 1 - \mathrm{pr}(T = M = 1|W) - \mathrm{pr}(T = M = 0|W) \\
&= 1 - \mathrm{pr}(M = 1|W)\mathrm{pr}(T = 1|W) - (1 - \mathrm{pr}(M = 1|W))(1 - \mathrm{pr}(T = 1|W)) \\
&= \mathrm{pr}(M = 1|W) + \mathrm{pr}(T = 1|W) - 2\mathrm{pr}(M = 1|W)\mathrm{pr}(T = 1|W) \\
&= \mathrm{pr}(M = 1|W) + \mathrm{pr}(T = 1|W)(2\mathrm{pr}(M = 0|W) - 1).
\end{aligned}$$

Therefore, if $2\text{pr}(M = 0|W) - 1 < 0$, i.e., $\pi(W) \equiv \text{pr}(M = 0|W) \leq 1/2$, the misassignment probability $\text{pr}(T \neq M|W)$ is minimized when $\text{pr}(T = 1|W) = 1$, that is, assigning the patient to the treatment $T = 1$. Similarly, if $\text{pr}(M = 0|W) > 1/2$, $\text{pr}(T \neq M|W)$ is minimized when $\text{pr}(T = 0|W) = 1$, that is, assigning the patient to the treatment $T = 0$.

Proof of Theorem 2

Let $f(W)$ denote the density function of W, and define $C = p_{01}\text{pr}(M = 0) + p_{10}\text{pr}(M = 1)$, $D_0 = p_{00} - p_{01}$, $D_1 = p_{11} - p_{10}$, $\omega = D_0 + D_1$ and $\delta = D_1/D_0$. It follows that

$$\text{pr}(Y = 1) = \sum_{j=0}^{1}\sum_{k=0}^{1} \text{pr}(M = j, T = k)p_{jk}$$

$$= C + D_0 \int \text{pr}(M = 0|W)\text{pr}(T = 0|W)f(W)dW$$

$$+D_1 \int \text{pr}(M = 1|W)\text{pr}(T = 1|W)f(W)dW$$

$$= C + D_0 \int (1 - \text{pr}(M = 1|W))(1 - \text{pr}(T = 1|W))f(W)dW$$

$$+D_1 \int \text{pr}(M = 1|W)\text{pr}(T = 1|W)f(W)dW$$

$$= C + \int [D_0 \{1 - \text{pr}(M = 1|W)\} + \text{pr}(T = 1|W)\{D_1 - \omega\text{pr}(M = 0|W)\}]f(W)dW.$$

As a result, when $\omega > 0$ which indicates a positive predictive marker effect, if $\pi(W) \equiv \text{pr}(M = 0|W) \leq D_1/\omega = \delta/(1 + \delta)$, $\text{pr}(Y = 1)$ is maximized when $\text{pr}(T = 1|W) = 1$, that is, assigning the patient to the treatment $T = 1$; and if $\pi(W) > \delta/(1 + \delta)$, $\text{pr}(Y = 1)$ is maximized when $\text{pr}(T = 1|W) = 0$, that is, assigning the patient to the treatment $T = 0$. Similarly, when $\omega \leq 0$ which indicates a negative predictive marker effect, if $\pi(W) \equiv \text{pr}(M = 0|W) \leq \delta/(1 + \delta)$, $\text{pr}(Y = 1)$ is maximized when assigning the patient to the treatment $T = 0$; and if $\pi(W) > \delta/(1 + \delta)$, $\text{pr}(Y = 1)$ is maximized when assigning the patient to the treatment $T = 1$. In general, if $\pi(W) \leq \delta/(1 + \delta)$, $\text{pr}(Y = 1)$ is maximized when assigning the patient to the treatment $T = \text{I}(\omega > 0)$; and otherwise to $T = 1 - \text{I}(\omega > 0)$.

References

Carroll, RJ., Ruppert, D., Stefanski, LA. and Crainiceanu, C. (2006), “Measurement error in nonlinear models: a modern perspective,” *CRC Press: London*.

Dempster AP., Laird NM., and Rubin DB. (1977), “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society, Series B* 39, 1–38.

Freidlin, B., Jiang W. and Simon R. (2010), “The cross-validation adaptive signature design,” *Clinical Cancer Research*, 16, 691–698.

Fuller, WA. (1987), “Measurement error models,” *Wiley: New York, NY*.

Ibrahim JG. (1990), “Incomplete data in generalized linear models,” *Journal of the American Statistical Association*, 85, 765–769.

Mandrekar, S.J., and Sargent, D.J. (2009), “Clinical Trial Designs for Predictive Biomarker Validation: Theoretical Considerations and Practical Challenges,” *Journal of Clinical Oncology*, 27, 4027–4034.

Sledge, GW. (2005), “What is target therapy,” *Journal of Clinical Oncology*, 23, 1614–1615.

Statistical Considerations for Evaluating Prognostic Biomarkers: Choosing Optimal Threshold

Zheng Zhang

Abstract The use of biomarker is increasingly popular in cancer research and various imaging biomarkers have been developed recently as prognostic markers. In practice, a threshold or cutpoint is required for dichotomizing continuous markers to distinguish patients with certain conditions or responses from those who are without. Two popular ROC based methods to establish "optimal" threshold are based on Youdan index J and closest top-left criterion. We have shown in this paper the importance to acknowledge the inherent variance of such estimates. In addition, a purely data-driven approach to search for optimal threshold can produce estimates that are not necessarily meaningful due to the large variance in such estimates. Instead, we propose to estimate the threshold through pre-specified criterion, such as a fixed level of specificity. The confidence intervals of the threshold and sensitivity at the pre-specified specificity are much narrower compared to the quantities measured through either Youdan index J or closest top left criterion. We suggest to estimate the threshold at a pre-specified level of specificity, and the sensitivity at that threshold, all the estimates should be accompanied by appropriate 95 % confidence intervals.

Keywords Biomarker • ROC • Threshold • Optimal • Youdan index

1 Introduction

From various clinical studies conducted during the past decade, a large collection of biomarkers have been studied on their abilities to predict important clinical outcomes such as treatment response, progression-free survival and overall survival in patients who were diagnosed with cancer and under treatment. One group of such markers have been derived from advanced imaging procedures, such as rCBV from dynamic susceptibility contrast-enhanced (DSC) MR perfusion (Paulson and Schmainda 2008), K^{trans} from dynamic contrast-enhanced (DCE) MR perfusion (Sourbron and Buckley 2013) and ADC values from diffusion-weighted

Z. Zhang (✉)
Department of Biostatistics, Center for Statistical Sciences, Brown University School of Public Health, Providence, RI 02912, USA

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_2

imaging (DWI) (Bihan et al. 2006). Those markers are usually measured at several time-points throughout the study, such as pre-treatment, mid-treatment and post-treatment. The most frequently used marker values are those either measured at pre-treatment, or changes in marker values from pre-treatment measurements to various after treatment measurements. Due to the continuous nature of those values, the clinical usefulness of such markers often depends on whether a threshold can be determined to classify the marker. For example, a marker value above that threshold would predict a favorable outcome (better response to treatment, longer survival, etc.) and a marker value below that threshold would predict an unfavorable outcome. For all the possible thresholds that can be found, we would want to determine whether there is an optimal threshold that offers the best predictive performance.

2 A Brief Review of the ROC Curve

The receiver operating characteristic (ROC) curve (Swets and Pickett 1982; Pepe 2003) is a popular statistical tool to define predictive accuracy, hence it provides a pathway to determine the optimal threshold. The ROC curve is a collection of pairs of sensitivities and specificities, each pair is determined by a unique threshold. Assuming a test is done to diagnose a disease, the ROC curve is a plot of sensitivity versus 1-specificity, where sensitivity is the probability of the test value to correctly identify disease and specificity is the probability of it to correctly identify non-disease cases. The ROC curve can be written as a function of $t \in (0, 1)$, by letting $\bar{D}$ and D denote non-diseased and diseased populations and $S_{\bar{D}}$ and S_D be the survivor functions for test result Y from $\bar{D}$ and D, respectively, such as $S_D(c) = P[Y \geq c|D]$, $S_{\bar{D}}(c) = P[Y \geq c|\bar{D}]$, then the ROC curve is defined as $ROC(t) = S_D(S_{\bar{D}}^{-1}(t)), \quad t \in (0, 1)$.

To estimate the ROC curve empirically from test results $Y = \{Y_{D,i}, Y_{\bar{D},j}\}$, $i = 1, \ldots, n_D, j = 1, \ldots, n_{\bar{D}}$, $N = n_D + n_{\bar{D}}$, define $\widehat{sen}(c) = \sum_{i=1}^{n_D} I[Y_{D_i} \geq c]/n_D$ and $\widehat{1 - spec}(c) = \sum_{j=1}^{n_{\bar{D}}} I[Y_{\bar{D}_j} \geq c]/n_{\bar{D}}$, then the empirical ROC curve is a plot of $\widehat{sen}(c)$ versus $\widehat{1 - spec}(c)$ for all possible cut points c on the real line.

The area under the ROC curve (AUC) is commonly used to determine the discrimination power of the test. It is defined as

$$AUC = P(Y_D > Y_{\bar{D}}) \tag{1}$$

The empirical AUC is estimated as a Mann-Whitney U-Statistics

$$\widehat{AUC} = \sum_{i=1}^{N} \sum_{j=1}^{N} \{I[Y_{D,i} > Y_{\bar{D},j}] + \frac{1}{2} I[Y_{D,i} = Y_{\bar{D},j}]\}/N^2 \tag{2}$$

3 Criteria Based on the ROC Curve

The criteria based on the ROC curve seek to maximize sensitivity and specificity simultaneously. Two such criteria are frequently used: The first one is called Youdan index J (Youdan 1950), which is the threshold corresponding to the point on the ROC curve that has the longest distance to the identity (diagonal) line. Hence this threshold is chosen to maximize the sum of sensitivity and specificity. Intuitively, this point is the point on the ROC curve that is the furthest away from the curve corresponds to a "useless" test. First define the distance from a point on the ROC curve to the diagonal line as D and c is the threshold corresponding to that point, then $D = \sqrt{(sen(c) + spec(c) - 1)^2/2}$ and Youdan index J is $J = sen(c) + spec(c) - 1$.

The second criterion, "closest top left" criterion (Perkins and Schisterman 2006) identifies the point on the ROC curve that had the shortest distance to the top-left corner (a point that confers the perfect test). This criterion seeks to minimize the sum of squares of false positive rate and false negative rate. Intuitively, this point is the point on the ROC curve that is closest to point with perfect sensitivity and perfect specificity. Here the distance $D = \sqrt{(1 - sen(c))^2 + (1 - spec(c))^2}$.

4 Issues When Reporting the Optimal Threshold

The optimal thresholds determined through either Youdan index J or "closest top left" criteria that were reported in the medical or statistical literature have seldom been accompanied by any measures of uncertainty. We should be aware that since either threshold is estimated from the ROC curve, there are inherent variances associated with the threshold estimates. This motivated our simulation studies to assess the variability in threshold estimation.

5 Simulation Study

We had simulated data from normal distribution with 100 or 200 subjects, evenly distributed between diseased and non-diseased subjects. The parameters of the normal distribution are chosen with AUC of 0.760 or 0.814. The ROC curve and its AUC are estimated empirically and the variabilities of the estimations are evaluated through 1000 bootstrap samples. We report empirical AUC, optimal thresholds and their associated sensitivities and specificities. For each quantity, we will calculate the exact 95 % bootstrap confidence intervals (CI).

We first generated the data as $Y_{\bar{D}} \sim N(0, 1)$ and $Y_D \sim N(1, 1)$ so that the true AUC is 0.760.

Table 1 shows the simulation results. For $N = 200$, we found empirical AUC to be 0.761(95 % CI: 0.693 to 0.827). The optimal threshold is 0.448(95 % CI:

Table 1 Thresholds and the associated accuracy measures

	Youdan	Top-left	Spec=0.70	Spec=0.90
N=(50,50), AUC=0.760				
Threshold	0.428(−0.232,1.104)	0.496(0.108,0.897)	0.512(0.176,0.869)	1.244(0.805,1.712)
Sensitivity	0.75(0.52,0.94)	0.72(0.58,0.86)	0.69(0.50,0.86)	0.41(0.20,0.64)
Specificity	0.70(0.44,0.92)	0.72(0.58,0.86)	–	–
N=(100,100), AUC=0.760				
Threshold	0.448(−0.112,1.005)	0.491(0.176,0.794)	0.517(0.263,0.786)	1.265(0.969,1.592)
Sensitivity	0.73(0.54,0.90)	0.71(0.60,0.82)	0.69(0.55,0.81)	0.40(0.24,0.56)
Specificity	0.70(0.50,0.87)	0.71(0.60,0.81)	–	–
N=(50,50), AUC=0.814				
Threshold	0.568(0.206,0.951)	0.445(0.187,0.714)	0.256(0.077,0.438)	0.627(0.404,0.877)
Sensitivity	0.70(0.52,0.86)	0.74(0.62,0.86)	0.77(0.64,0.88)	0.65(0.48,0.80)
Specificity	0.89(0.72,1.00)	0.83(0.70,0.94)	–	–
N=(100,100), AUC=0.814				
Threshold	0.593(0.300,0.893)	0.445(0.251,0.634)	0.259(0.123,0.392)	0.633(0.475,0.805)
Sensitivity	0.68(0.54,0.80)	0.73(0.63,0.81)	0.77(0.67,0.86)	0.64(0.53,0.73)
Specificity	0.89(0.77,0.98)	0.83(0.73,0.91)	–	–

−0.112 to 1.005) using Youdan's index and 0.491(95 % CI: 0.176 to 0.794) using the closest top left criterion. The estimated sensitivity is 0.73(95 % CI: 0.54 to 0.90) or 0.71(95 % CI: 0.60 to 0.82) and the estimated specificity is 0.70(95 % CI: 0.50 to 0.87) or 0.71(95 % CI: 0.60 to 0.81), respectively.

We next simulated data as $Y_{\bar{D}} \sim N(0, 0.5)$ and $Y_D \sim N(1, 1)$ so that the true AUC is 0.814. For N=200, the empirical AUC was estimated to be 0.813(95 % CI: 0.743 to 0.872). The optimal threshold is 0.593(95 % CI: 0.300 to 0.893) using Youdan's index and 0.445(95 % CI: 0.251 to 0.634) using the closest top left criterion. The estimated sensitivity is 0.68(95 % CI: 0.54 to 0.80) or 0.73(95 % CI: 0.63 to 0.81) and the estimated specificity is 0.89(95 % CI: 0.77 to 0.98) or 0.83(95 % CI: 0.73 to 0.91), respectively.

Optimal threshold based on the Youdan index tends to have wider confidence intervals than the threshold estimated through the top-left corner criterion. Compared to the same quantities estimated from the top left corner criterion, the associated sensitivity at the Youdan's threshold is lower, but the associated specificity is higher, and both have wider confidence intervals.

However, the utility of "optimal threshold" is debatable. As shown above, the optimal thresholds and their associated sensitivities and specificities all have large variance and are hard to interpret. We instead propose to estimate the threshold corresponding to a pre-specified criterion, such as a fixed specificity. As shown in Table 1, we had estimated he threshold values corresponding to the fixed specificity level of 70 % or 90 %, and the associated sensitivities at those thresholds. For N=200 and AUC=0.814, the threshold is 0.259(95 % CI 0.123 to 0.392) at 70 %

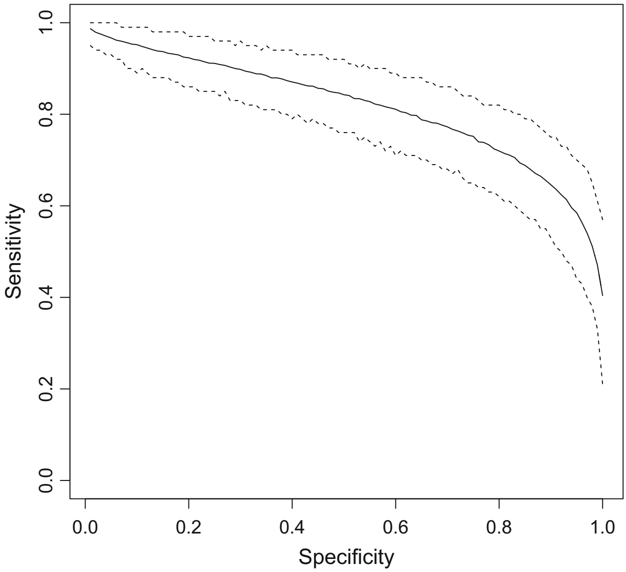


Fig. 1 Sensitivity-Specificity plot. The estimated sensitivities at the fixed specificity levels with the 95 % confidence intervals are shown

specificity and the associated sensitivity is 0.77(95 % CI 0.67 to 0.86). For 90 % specificity, the threshold is 0.633(95 % CI 0.475 to 0.805), the associated sensitivity is 0.64(95 % CI 0.53 to 0.73).

We had further estimated the thresholds and the associated sensitivities within the range of specificities of 0–99 % for N=200 and AUC=0.814, and plotted the estimated sensitivities versus the specificities in Fig. 1, which we named as a sensitivity-specificity plot, and included the point-wise 95 % confidence intervals for the estimated sensitivities. Similarly, Fig. 2 is a threshold-specificity plot, which presented both the estimated thresholds and the corresponding point-wise 95 % confidence intervals. Using both figures, we wanted to demonstrate the approach of finding the thresholds and the associated performance matrix for a prognostic biomarker.

6 Discussion

In estimating the optimal threshold, it is important to acknowledge the inherent variance of such estimates. In addition, a purely data-driven approach to search for optimal threshold can produce estimates that are not necessarily meaningful due to the large variance in such estimates. Instead, we propose to estimate the threshold through pre-specified criterion, such as a fixed level of specificity. The confidence intervals of the threshold and sensitivity at the pre-specified specificity are much

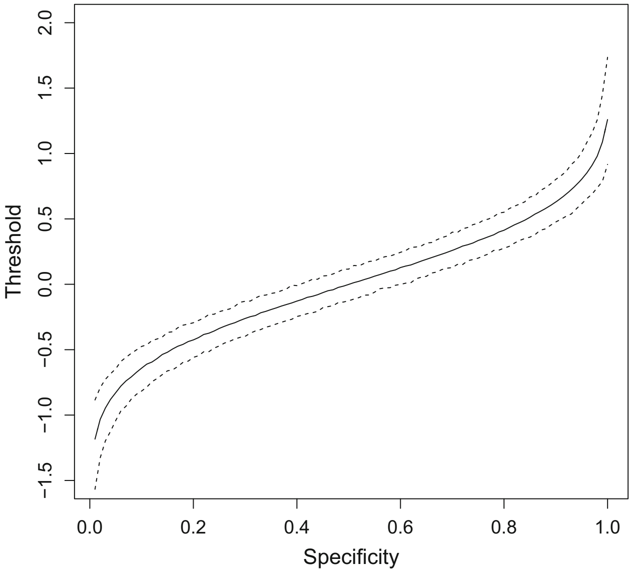


Fig. 2 Threshold-Specificity plot. The estimated thresholds at the fixed specificity levels with the 95 % confidence intervals are shown

narrower compared to the quantities measured through either Youdan index or closest top left criterion. We hereby suggest to estimate the threshold at a pre-specified level of specificity, and then estimate the sensitivity at that threshold, and all the estimates should be accompanied by appropriate 95 % confidence intervals.

References

Bihan, D. L., Urayama, S., Aso, T., Hanakawa, T. and Fukuyama, H. (2006). Direct and fast detection of neuronal activation in the human brain with diffusion MRI. *PNAS*, 103(21): 8263–8268.

Paulson, E. S. and Schmainda, K. M. (2008). Comparison of Dynamic Susceptibility-weighted Contrast-enhanced MR method: recommendations for measuring relative cerebral blood volume in brain tumors. *Radiology*, 249(2): 601–613.

Pepe, M. S. (2003). The statistical evaluation of medical tests for classification and prediction. Oxford University Press, United Kingdom.

Perkins, N. J. and Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am. J. Epidemiol.*, 163(7): 670–675.

Sourbron, S. P. and Buckley, D. L. (2013). Classic models for dynamic contrast-enhanced MRI. *NMR Biomed.*, 26: 1004–1027.

Swets, J. A. and Pickett, R. M. (1982). Evaluation of diagnostic systems: method from signal detection theory. Academic Press.

Youdan, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3: 32–35.

Accuracy of Meta-Analysis Using Different Levels of Diagnostic Accuracy Measures

Yanyan Song, Ying Lu, and Lu Tian

Abstract Diagnostic studies report results in sensitivity and specificity, or figures of receiver operation characteristics (ROC) curve. Meta-analysis synthesizes these diagnostic accuracy measures from different studies to obtain an overall summary ROC curve. Increasingly, meta analysis also uses individual patient level data. However, the pro and con of such an approach are not entirely clear. In this paper, we performed a simulation study to evaluate the accuracy of summary ROC curves derived from different types of data, i.e., the paired sensitivity and specificity from individual study, the study-specific ROC curves, and the individual patient level data. Extensive simulation experiments were conducted under various settings to compare the empirical performance of estimated summary ROC curves using data from three levels. The simulation results demonstrated that the method based on reported ROC curves from individual study provides accurate and robust summary ROC curve compared with alternatives including those based on patient level data and is preferred in practice.

Y. Song
Department of Biomedical Data Science, Stanford University School of Medicine,
HRP Redwood Building, Stanford, CA 94305-5405, USA

Department of Pharmacology and Biostatistics, Institute of Medical Sciences,
Shanghai Jiao Tong University, Shanghai, China

Y. Lu
Department of Biomedical Data Science, Stanford University School of Medicine,
HRP Redwood Building, Stanford, CA 94305-5405, USA

VA Palo Alto Health Care System, Palo Alto, CA, USA

L. Tian (✉)
Department of Biomedical Data Science, Stanford University School of Medicine,
HRP Redwood Building, Stanford, CA 94305-5405, USA

Department of Statistics, Stanford University, Stanford, CA, USA
e-mail: lutian@stanford.edu

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_3

Keywords Sensitivity • Specificity • Receiver operating characteristics (ROC) curve • Meta-analysis • Bivariate normal random effects model • Simulation experiment

1 Introduction

Diagnostic and predictive tests are important components of medical care. Clinicians rely on test results to establish diagnosis and guide patient management (Abbas et al. 2007). For example, diagnostic tests using either longitudinal Magnetic Resonance Imaging (MRI) or X-ray scans of hands can measure the total joint erosion and its change overtime to monitor the disease progression of rheumatoid arthritis (RA). A clinical decision on whether to change treatments often depends on the MRI or X-ray test results on RA progression status. MRI, a 3-dimensional technology that provides a better view of bone structure than a 2-dimensional X-ray, is anticipated to be able to provide an earlier and more sensitive prediction of disease progression.

Comparing results of a diagnostic test with a reference standard (in our RA example, clinical progression in 60 months), also known as the gold standard in clinical practice, results in a simple 2×2 table. In general, the gold standard is the presence or absence of a target condition, which may be based on a combination of tests (Pepe 2004). Sensitivity and specificity, as well as positive and negative predictive values can be calculated as measurements of the diagnostic accuracy (Zhou et al. 2014; Naaktgeboren et al. 2013). These diagnostic accuracy measures tell us about the ability of a diagnostic test to discriminate between and/or predict the presence and absence of the target condition, often referred to as disease and healthy status. An ideal diagnostic test should have no misclassification errors (false negatives or false positives). However, a perfect diagnostic test almost has never existed in clinical practice. Accuracy measures allow us to quantitatively compare several imperfect diagnostic tests.

For a given diagnostic test and target population, the accuracy measures in sensitivity and specificity can vary from different studies depending on the aggressiveness of decision makers. For a high risk RA population, one may want to aggressively call disease progression with a small amount of increase in the total erosion score in order to increase the test sensitivity. An increase in sensitivity will result in an increase in positive predictive value at the expenses of a decrease in specificity and the negative predictive value (Swets 1982). Thus, accuracy measures in sensitivity and specificity, or equivalently expressed as the true positive rate (TPR) and false positive rates (FPR = 1-specificity), are difficult to compare directly between technologies or studies. An alternative measurement of diagnostic accuracy is the receiver operating characteristics (ROC) curve that is independent of the disease prevalence and selection of the binary decision threshold for a positive conclusion of a diagnostic test (Metz 1986; Hanley 1989). The statistical interpretation of the area under an ROC curve (AUC) is the probability of making a correct diagnosis for a given pair of a disease and a healthy subject (Hanley and McNeil 1982). A higher AUC implies a higher accuracy of a diagnostic test.

In medical literature, diagnostic accuracies are often reported in sensitivities and specificities (Level 1). Increasingly, ROC curves (Level 2) are presented graphically and AUCs under an ROC curves are reported. In rare occasions, the individual patient level data used to estimate the ROC curve may also be available (Level 3).

Most diagnostic studies have small or moderate sample sizes that may be insufficient for accurate estimates for the diagnostic precision (Bachmann et al. 2006). Studies may also vary in eligibility criterion that attribute to variations in accuracy measurements and affect generalizability of results from individual study. The meta-analytic approach is often required to overcome these limitations by pooling studies to evaluate the same diagnostic technology (Zhou et al. 2015) and provide a more precise estimate of test performance than that from a single study (Pai et al. 2004; Knottnerus et al. 2002).

Meta-analysis for diagnostic test accuracy (DTA) has many methodological challenges because of the nature of observational study design (Leeflang et al. 2008; Leeflang et al. 2013). Sensitivity and specificity are likely to be correlated (between studies) due to threshold effects so that it is necessary to use multivariate analytic methods (Adams et al. 2009). The bivariate approach analyzes the sensitivity and specificity jointly (Reitsma et al. 2005). Hierarchical summary ROC (HSROC) model could also combine estimated pairs of sensitivity and specificity from multiple studies by extending the Moses-Littenberg fixed-effects summary ROC (SROC) model (Rutter and Gatsonis 2001). The HSROC approach can be readily used for conducting meta-analysis with the Level 1 data. With the availability of ROC curves reported in literature, it is possible to directly synthesize ROC curves by digital scans. Directly working with the resulting ROC curves has the advantage of having the entire range of accuracy measures of a diagnostic test. A summary ROC curve can then be derived for the Level 2 ROC curves. The most comprehensive approach for a meta-analysis is to use individual patient level data from all studies (Level 3 data). The advantages of having Level 3 data include more accurate estimation, possible verification of model assumptions and a better evaluation of study heterogeneity and effect of covariates. However, oftentimes individual patient data are not directly available. Even global regulatory agencies are publishing policies to require sharing clinical trial data on patient level, access to such data still faces many challenges and is limited, perhaps due to the concerns for patient privacy and risk for individual identification of study participants (Koenig 2999).

Since all three nested levels of data can be used in meta-analyses to derive summary ROC curves for a diagnostic test, a question is whether these resulting ROC curves have similar accuracy. If reporting sensitivity and specificity is not adequate, we may want to require plots of ROC curves in publications or even individual patient data. On the other hand, if the access to individual level data is not necessary, it will substantially reduce the workload for performing a good meta-analysis. This paper is to answer this question via a simulation study. Sect. 2 presents study design: firstly, accuracy measures and model parameters are described; then the statistical methods to synthesize ROC curves using data from different levels are presented; and finally, the experimental design for our simulation study is presented. Sect. 3 reports the results of the simulation study. Sect. 4 concludes the study with further discussions.

2 Methods

2.1 Definitions of Diagnostic Accuracy Measures

Let Y be a random variable for the binary disease status and X be the continuous random variable of result of a diagnostic test. A high value of X means an increasing risk for having the disease. If X being greater than c is defined as a positive test, the *FPR* and *TPR* are defined as

$$\begin{aligned} FPR &= Pr\left(X > c \middle| Y = 0\right) = S_0(c) \\ TPR &= Pr\left(X > c \middle| Y = 1\right) = S_1(c) \end{aligned} \tag{1}$$

where $S_i(.)$ is the survival distribution function for X in subpopulations $Y = i$. It is clear that both *FPR* and *TPR* are decreasing functions of c. A ROC curve represents the *TPR* as a function of *FPR*, thus, links these two diagnostic accuracy measures. Mathematically,

$$ROC(u) = S_1\left\{S_0^{-1}(u)\right\} \tag{2}$$

for u in between [0,1].

2.2 A Bivariate Normal Model for Data Generation for Simulation Studies

A bivariate normal model assumes that the underlying diagnostic variable for a given study follows two normal distributions $N(\mu_0, \sigma_0^2)$ and $N(\mu_1, \sigma_1^2)$ for healthy and disease populations, respectively. In this case, $S_i(u) = 1 - \Phi\left(\frac{u-\mu_i}{\sigma_i}\right)$ for $i = 0$, 1, where $\Phi(\bullet)$ is the cumulative distribution of the standard normal distribution. Although a bivariate normal model may appear to be restrictive, it can fit for quite general conditions (Metz and Pan 1999), thus is the most popular parametric approach for ROC analysis in practice.

Let k indicates the kth study in a meta-analysis. For the kth study, the diagnostic variable $X_{k,i}$ for the ith subject follows the bivariate normal model:

$$X_{k,i} = \mu_0\left(1 - Y_{k,i}\right) + \mu_1 Y_{k,i} + \left\{\sigma_0\left(1 - Y_{k,i}\right) + \sigma_1 Y_{k,i}\right\}\varepsilon_{k,i} + e_{k0}\left(1 - Y_{k,i}\right) + e_{k1} Y_{k,i} \tag{3}$$

Table 1 Bivariate normal model parameters for simulation studies

μ_0	μ_1	σ_0	σ_1	σ_{00}	σ_{11}	σ_{01}
0.34	0.71	0.298	0.364	0.010	0.020	0.014

Here, $\varepsilon_{ki} \sim N\,(0, 1)$. The difference from a conventional bivariate normal model is the inclusion of study level random effects in meta-analysis, which is characterized by the random vector $\begin{pmatrix} e_{k,0} \\ e_{k,1} \end{pmatrix} \sim N \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right\}$, where $\varSigma = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01}^2 \\ \sigma_{01}^2 & \sigma_{11}^2 \end{bmatrix}$. Thus, even the diagnostic values of cases and controls are independent within a study, they are correlated in the meta-analysis. Under this random effects model, the underlying population ROC curve to be estimated is then the expected ROC curve over random effects, i.e.,

$$ROC_C(u) = E\left\{ROC_k\left(u, e_{k1,} e_{k0}\right)\right\},$$

where $ROC_k\left(u, e_{k1}, e_{k0}\right) = \Phi\left\{\frac{\mu_1 - \mu_0 + e_{k1} - e_{k0}}{\sigma_1} + \frac{\sigma_0}{\sigma_1}\Phi^{-1}(u)\right\}$.

In this paper, we will not use the data from our motivating RA example due to confidentiality agreement. However, we use the estimated parameters based on the real data for model simulation study. The model parameters used in our simulation are specified in Table 1.

Additional model parameters to be specified for our simulation study are sample sizes of individual studies and number of studies. Let n_k , n_{0k} , and n_{1k} denote the total sample size, the number of controls and cases, respectively, in the kth study. In our simulation, we randomly drew $\{n_1, \ldots, n_K\}$ from $\{60, 64, \ldots, 196, 200\}$ and set $n_{1k}/n_k = 25\,\%$ and $K = 8$ or 20.

2.3 Meta-Analysis for Level 1 Data

For meta-analysis using level 1 data, we assume that only one pair of FPR and TPR are reported from a study. In the simulation study, for the kth study, we drew a random FPR_k from the uniform distribution $U\,(0.1, 0.5)$, which specified the cutoff value c_k and the corresponding TPR: (FPR_k, TPR_k). Noting that

$$\begin{aligned} \Phi^{-1}\left(1 - FPR_k\right) &= \frac{c_k - \mu_{k0}}{\sigma_0} \\ \Phi^{-1}\left(1 - TPR_k\right) &= \frac{c_k - \mu_{k1}}{\sigma_1} \end{aligned} \tag{4}$$

we have

$$\sigma_0 \Phi^{-1}\left(1 - FPR_k\right) + \mu_{k0} = \sigma_1 \Phi^{-1}\left(1 - TPR_k\right) + \mu_{k1}.$$

Let $\psi_k = \Phi^{-1}(1 - TPR_k)$ and $\xi_k = \Phi^{-1}(1 - FPR_k)$, their relationship can be characterized as

$$\psi_k = \frac{\mu_0 - \mu_1}{\sigma_1} + \frac{\sigma_0}{\sigma_1}\xi_k = \alpha + \beta\xi_k \tag{5}$$

Based on $\{(\psi_k, \zeta_k), k = 1, \ldots, K\}$, we may obtain the least square estimator for α and β and the ROC curve can be summarized as

$$\widehat{ROC}_A(u) = \Phi\left(-\widehat{\alpha} + \widehat{\beta}\,\Phi^{-1}(u)\right) \tag{6}$$

2.4 Meta-Analysis for Level 2 Data

As we explained previously in Sect. 1, AUC of an ROC is simply the probability of correct diagnosis for a pair of disease and healthy subjects. A simple way to perform the meta-analysis is via a weighted linear combination of the study-specific ROC curves based on $ROC_k(u)$ for u in the interval [0,1] and $k = 1, \ldots, K$:

$$\widehat{ROC}_B(u) = \sum_{k=1}^{K} w_k(u)\widehat{ROC}_k(u) \tag{7}$$

where $\sum_{k=1}^{K} w_k(u) = 1$.

In the fixed effect model, the optimal weight is equal to the inverse of the estimated variance, which is proportional to

$$w_k(u) \propto \left\{\text{var}\left(\widehat{ROC}_k(u)\right)\right\}^{-1} = \left\{\frac{\widehat{ROC}_k(u)\left[1 - \widehat{ROC}_k(u)\right]}{n_{k1}} + \frac{f_{k1}(c_k)^2}{f_{k0}(c_k)^2}\frac{u\,(1-u)}{n_{k0}}\right\}^{-1} \tag{8}$$

Here $f_{k1}(\cdot)$ and $f_{k0}(\cdot)$ are the density functions of the biomarker in cases and controls from study k, respectively. In a random effect model, the optimal weight is equal to

$$w_k(u) \propto \left\{\text{var}\left(\widehat{ROC}_k(u)\right) + \tau(u)\right\}^{-1} \tag{9}$$

where $\tau(u)$ is the variance of the random effect across studies. Therefore, to use the aforementioned weights in Eq. (9), unknown quantities in the weights need to be estimated. A simple choice in practice, however, is to let

$$w_k(u) = \frac{n_{k0}n_{k1}}{n_k} \Bigg/ \sum_{l=1}^{K} \frac{n_{l0}n_{l1}}{n_l} \tag{10}$$

which still derives a consistent estimate of $ROC_B(u)$. We used the simple weight (10) in our simulation study.

2.5 Meta-Analysis for Level 3 Data

With individual level data $(X_{ki},\ Y_{ki}\ )$ for$i = 1,\ldots,n_k$ for study $k = 1,\ldots,K$, we can fit a mixed effect regression model based on Eq. (3). The summary population ROC curve is then can be the estimated expectation

$$ROC_C(u) = E\left\{\widehat{ROC}_k\left(u, e_{k1,}e_{k0}\right)\right\} \tag{11}$$

where the estimated model parameters for given random effects,

$$\widehat{ROC}_k\left(u, e_{k1}, e_{k0}\right) = \Phi\left\{\frac{\widehat{\mu}_1 - \widehat{\mu}_0 + e_{k1} - e_{k0}}{\widehat{\sigma}_1} + \frac{\widehat{\sigma}_0}{\widehat{\sigma}_1}\Phi^{-1}(u)\right\} \tag{12}$$

and $(e_{k0}, e_{k1})'$ satisfies with a bivariate normal distribution that can be estimated as $N\left\{(0,0)', \widehat{\Sigma}\right\}$.

We used a two-step procedure to fit the mixed effect regression model. First, we establish two separate mixed effect models as below:

$$\begin{aligned} X_{k,i}\,|(Y_{k,i} = 1) &= \mu_1 + e_{k1} + \sigma_1\varepsilon_{k,i} \\ X_{k,i}\,|(Y_{k,i} = 0) &= \mu_0 + e_{k0} + \sigma_0\varepsilon_{k,i} \end{aligned} \tag{13}$$

to obtain consistent estimators for σ_0 and σ_1, denoted as $\widehat{\sigma}_0$ and $\widehat{\sigma}_1$, respectively.

Secondly, we transform the biomarker value as $X^*_{k,i} = X_{k,i}\left(1 - Y_{ki}\right) + X_{k,i}\widehat{\sigma}_0/\widehat{\sigma}_1 Y_{ki}$ and fitted the regular mixed effect model below

$$X^*_{k,i} = \tilde{\mu}_0\left(1 - Y_{k,i}\right) + \tilde{\mu}_1 Y_{k,i} + \tilde{e}_{k0}\left(1 - Y_{k,i}\right) + \tilde{e}_{k1}Y_{k,i} + \varepsilon_{k,i} \tag{14}$$

to obtain the estimates $\widehat{\tilde{\mu}}_0, \widehat{\tilde{\mu}}_1, \widehat{\tilde{\sigma}}_{00}, \widehat{\tilde{\sigma}}_{11}$, and $\widehat{\tilde{\sigma}}_{10}$ for model parameters $\tilde{\mu}_0, \tilde{\mu}_1, \tilde{\sigma}_{00}, \tilde{\sigma}_{11}$, and $\tilde{\sigma}_{10}$, respectively. Finally, we let $\widehat{\mu}_0 = \widehat{\tilde{\mu}}_0, \widehat{\mu}_1 = \widehat{\tilde{\mu}}_1\widehat{\sigma}_1/\widehat{\sigma}_0, \widehat{\sigma}_{11} = \widehat{\tilde{\sigma}}_{11}, \widehat{\sigma}_{10} = \widehat{\tilde{\sigma}}_{10}\widehat{\sigma}_1/\widehat{\sigma}_0$ and $\widehat{\sigma}_{00} = \widehat{\tilde{\sigma}}_{00}\widehat{\sigma}_1^2/\widehat{\sigma}_0^2$ and use the estimated covariance of random effects $\widehat{\Sigma} = \begin{bmatrix} \widehat{\sigma}_{00}^2 & \widehat{\sigma}_{01}^2 \\ \widehat{\sigma}_{01}^2 & \widehat{\sigma}_{11}^2 \end{bmatrix}$ to calculate

$$\widehat{ROC}_C(u) = E\left\{\widehat{ROC}_k\left(u, e_{k1}, e_{k0}\right)\middle|\widehat{\Sigma}\right\} \tag{15}$$

by numerical integration.

2.6 Simulation Experiments

Simulation studies were used to investigate the consistency and accuracy of the estimated ROC curves. In addition, we wanted to evaluate the consistency and accuracy under three settings. Experiment 1, a meta-analysis had 20 studies. Experiment 2, a meta-analysis had 8 studies. In both experiments 1 and 2, the data were generated under the bivariate normal model specified in Eq. (3).

The Experiment 3 examined the effect of model misspecification with again 20 studies used in a meta-analysis. However, the data were generated by the following log-normal random effects model:

$$\begin{aligned}\log\left(X_{k,i}\right) = \mu_0\left(1 - Y_{k,i}\right) + \mu_1 Y_{k,i} + \left\{\sigma_0\left(1 - Y_{k,i}\right) + \sigma_1 Y_{k,i}\right\}\varepsilon_{k,i} \\ + e_{k0}\left(1 - Y_{k,i}\right) + e_{k1} Y_{k,i}\end{aligned} \tag{16}$$

All analyses for Experiment 3 did not consider log-transformation, thus fitted a misspecified model.

In each Experiment setting, we generated 1000 sets of simulated data. Each set yield estimated ROC curves $ROC_A(u)$, $ROC_B(u)$ and $ROC_C(u)$ defined in (6), (7) and (11).

We define the estimation bias as $B(u) = E\left\{\widehat{ROC}(u) - ROC(u)\right\}$ and the mean of absolute bias (MAD) is defined as $MAD(u) = E\left|\widehat{ROC}(u) - ROC(u)\right|$, where $ROC(u)$ is the true ROC curve.

3 Simulation Results

Figures 1 and 2 summarize the results for Experiment 1. The estimated ROC curves using all three level data had bias close to 0, indicating the consistency of three approaches regardless the level of data used. However the MADs shown in Fig. 2 were not the same for different ROC estimates. The MAD was largest for Level 1 data and smallest for the Level 3 data. However, numerically, the differences in MAD are small.

Figures 3 and 4 summarize the results for Experiment 2. Similar to Experiment 1, the estimated ROC curves using all three level data had bias close to 0 and MADs had the same order for level of data. Only MADs were slightly higher than those in the Experiment 1 but also almost identical for all three approaches.

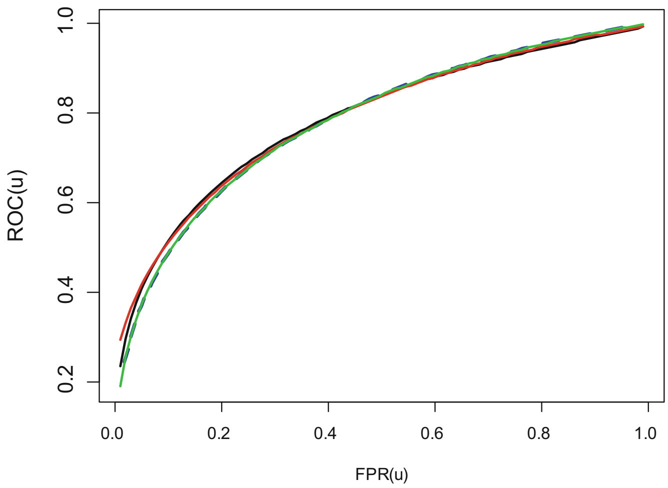


Fig. 1 Bias for the estimated ROC curves in experiment 1 Black (Level 1), Red (Leve 2), Green (Level 3) and Blue (True)

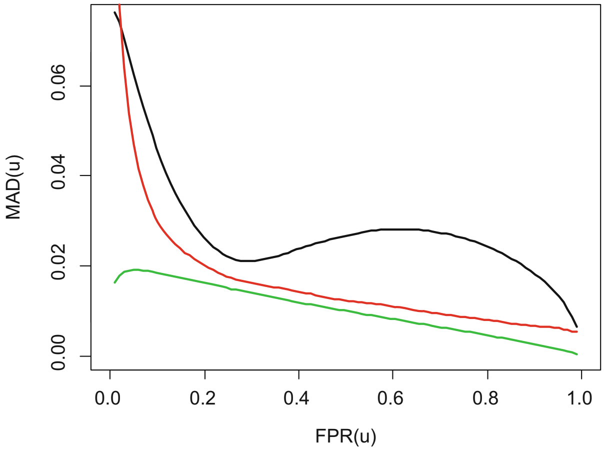


Fig. 2 Accuracy for the estimated ROC curves in experiment 1 Black (Level 1), Red (Leve 2), and Green (Level 3)

When data did not follow bivariate normal distribution in the Experiment 3, the results in Figs. 5 and 6 were different from the previous two experiments. As in Fig. 5, all estimated ROC curves had large bias. The MAD in Fig. 6 was the worst for method based on the Level 3 data and the smallest for the method based on the Level 2 data.

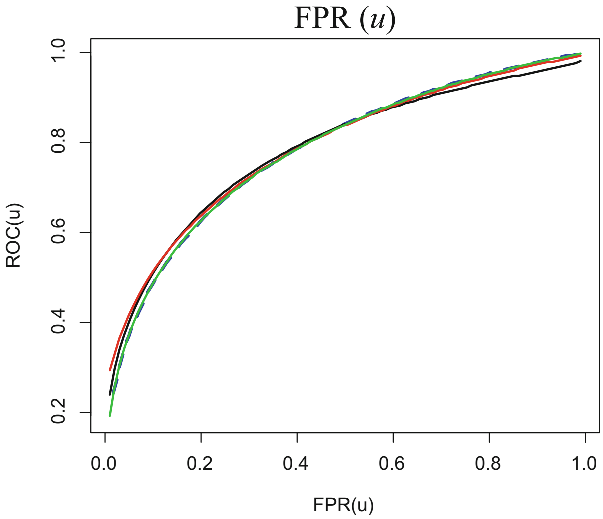


Fig. 3 Bias for the estimated ROC curves in experiment 2 Black (Level 1), Red (Leve 2), Green (Level 3) and Blue (True)

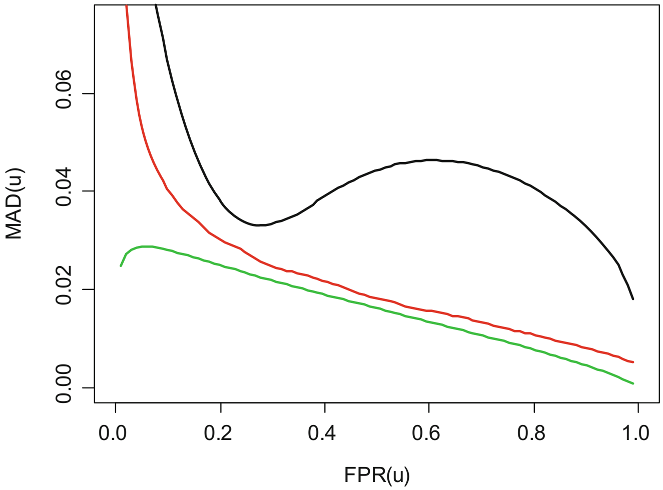


Fig. 4 Accuracy for the estimated ROC curves in experiment 2 Black (Level 1), Red (Leve 2), and Green (Level 3)

4 Discussion

In this paper, we performed meta-analysis for diagnostic tests using three different levels of information from individual studies. When the parametric model is correctly specified, all three levels of data provide almost unbiased estimator even

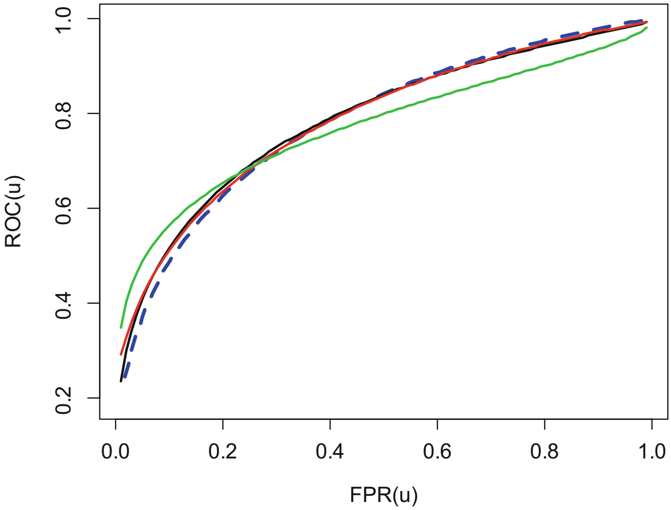


Fig. 5 Bias for the estimated ROC curves in experiment 3 Black (Level 1), Red (Leve 2), Green (Level 3) and Blue (True)

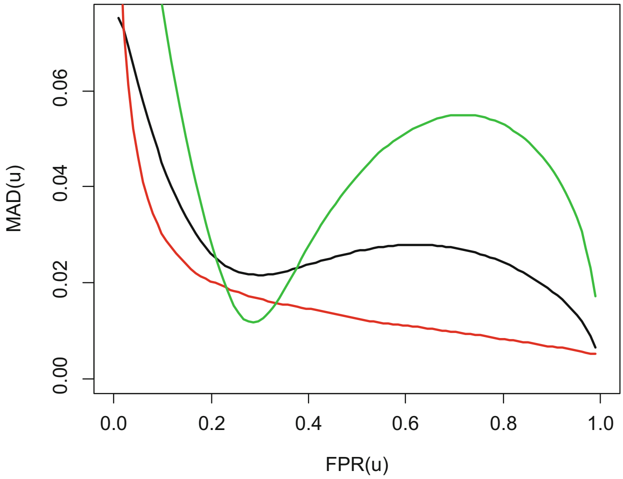


Fig. 6 Accuracy for the estimated ROC curves in experiment 3 Black (Level 1), Red (Leve 2), and Green (Level 3)

when the number of study is relatively small. When the parametric model is correctly specified, the individual level data can be used to improve the estimation accuracy. Especially, the gain against the analysis using pairs of FPR and TPR at Level 1 data can be substantial by using more information from individual studies, i.e., Level 2 and 3 data. The method based on Level 2 data with reported ROC curves from individual study, however, performed similarly to that based on the

Level 3 data with individual level information. When the parametric model is misspecified, using individual data can introduce systematic biases more than the other two methods in estimating the ROC curve. Overall, the Level 2 data provide a good compromise in both accuracy and robustness between Level 1 and Level 3 data and seems to be a preferred choice in practice.

Our simulation results have intuitive interpretations. A meta-analysis for Level 1 data transforms observed sensitivity and specificity to a binormal model for the ROC curve analysis. This approach is known to be robust against model misspecification (Metz and Pan 1999). Because it only uses one point on a ROC curve from each study, it is also least efficient. A meta-analysis for Level 2 data calculates a weighted average of observed ROC curves. While the calculation of the optimal weight depends on random effects model specification, the weighted average is always consistent. Using the entire ROC curve from each study improves its efficiency from only using one point in Level 1 data. The meta-analysis of Level 3 data utilizes individual patient level data to fit the summary ROC curve under the random effects model. Thus, it has the best efficiency and accuracy when the model is correctly specified, but worst accuracy otherwise.

Oftentimes, the objective of the meta-analysis is not only to summarize the collection of individual ROC curves, but also improve the estimation of individual ROC curve by borrowing information from other studies. To this end, more comprehensive statistical modelling such as the random effects model for Level 3 data is needed. It is beyond the scope of the current paper and warrants further research.

With the growing number of meta-analyses of test accuracy studies, we no longer question whether a meta-analysis of medical test accuracy studies is useful, rather what kind of data and methods are most suitable for such undertaking (Bachmann et al. 2004). This simulation study suggests that we should encourage diagnostic test accuracy studies to present high resolution plots of ROC curves in their research papers, in addition to commonly reported sensitivity and specificity pairs at given cut-off values. With the availability of ROC curves, it seems that the extra effort to request and get access of individual patient level data can be avoided without major loss of information for summarizing ROC curve.

The study has several limitations. First, the focus of our study is based on a summary measure of ROC curve. In addition to ROC curves, metrics for diagnostic accuracy can be measure by the summary odds ratio (Alvarez et al. 2006) and likelihood ratios (Abulafia and Sherer 1999). The impact of different levels of data on these summary measures was not evaluated in our simulation study.

Acknowledgements Y Song's work was supported by the National Natural Science Foundation of China (No. 81373395) and Shanghai Municipal Science and Technology Commissions (No. 14431905600). Y Lu's work was supported by US VA Cooperative Studies Program, USA. L Tian's work was partially supported by R01 HL089778-05 from NIH of USA.

References

Abbas SM, Bissett IP, Parry BR. Meta-analysis of oral water-soluble contrast agent in the management of adhesive small bowel obstruction. Br J Surg. 2007;94(4):404–11.

Pepe MS. (2004) The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, 2004. Oxford statistical science series, ISSN 0952–9942.

Zhou, X. H., Obuchowski, N. A. and McClish, D. K. (2014). Statistical Methods in Diagnostic Medicine. 2nd Ed. John Wiley and Sons, New York, NY. 45.

Naaktgeboren CA, Bertens LC, van Smeden M, et al. Value of composite reference standards in diagnostic research. BMJ 2013;347:f5605.

Swets JA. Sensitivities and specificities of diagnostic tests. JAMA 1982;248 (5):548–9.

Metz CE. ROC methodology in radiology imaging. Invest Radiol 1986;21:720–33.

Hanley J. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagnost Imag. 1989;29:307–35.

Hanley J, McNeil B. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.

Bachmann LM, Puhan MA, ter Riet G, et al. Sample sizes of studies on diagnostic accuracy: literature survey. BMJ 2006;332:1127–9.

Zhou X, Fang J, Yu C, Xu Z, Tian L, and Lu Y. (2015) Chapter 9, Meta-Analysis. Advanced Medical Statistics 2nd Ed. (Lu Y, Fang J, Tian L, and Jin H, editors), World Scientific Publishing Co. Pte. Ltd., New Jersey, USA, 2015.

Pai M, McCulloch, Enanoria W, Colford JM. Systematic reviews of diagnostic test evaluations: What's behind the scenes? ACP Journal Club. 2004; 141: A-11.

Knottnerus JA, ed. The evidence base of clinical diagnosis. London, UK: BMJ Books, 2002

Leeflang MM, Deeks JJ, Gatsonis C, et al., Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008;149:889–97.

Leeflang MM, Deeks JJ, Takwoingi Y, et al. Cochrane diagnostic test accuracy reviews. Syst Rev 2013;2:82.

Adams K, Shah PL, Edmonds L, et al. Test performance of endobronchial ultrasound and transbronchial needle aspiration biopsy for mediastinal staging in patients with lung cancer: systematic review and meta-analysis. Thorax. 2009;64(9):757–62.

Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology 2005; 58(10):982–990.

Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001; 20(19):2865–2884.

Koenig F, Slattery J, Goves T, Lang T, Benjamini Y, Day S, Bauer P, and Posch M. Sharing clinical trial data on patient level: Opportunities and challenges. Biometrical Journal 57(2015)1, pp 8–26.

Metz CE and Pan X. "Proper" Binormal ROC Curves: Theory and Maximum-Likelihood Estimation. Journal of Mathematical Psychology 1999; 43, 1–33.

Bachmann LM, Haberzeth S, Steurer J, et al. The accuracy of the Ottawa knee rule to rule out knee fractures: a systematic review. Ann Intern Med. 2004;140(2):121–4.

Alvarez S, Anorbe E, Alcorta P, et al. Role of sonography in the diagnosis of axillary lymph node metastases in breast cancer: a systematic review. AJR Am J Roentgenol. 2006;186(5):1342–8.

Abulafia O, Sherer DM. Automated cervical cytology: meta-analyses of the performance of the Auto Pap 00 QC System. Obstet Gynecol Surv. 1999;54(7):469–76.

Part II
Bayesian Methods and Applications

Bayesian Frailty Models for Multi-State Survival Data

Mário de Castro, Ming-Hui Chen, and Yuanye Zhang

Abstract Multi-state models can be viewed as generalizations of both the standard and competing risks models for survival data. Models for multi-state data have been the theme of many recent published works. Motivated by bone marrow transplant data, we develop a Bayesian model using the gap times between two successive events in a path of events experienced by a subject. Path specific frailties are introduced to capture the dependence among the gap times sharing the same path with two or more states. In this study, we focus on a single terminal event. Under improper prior distributions for the parameters, we establish propriety of the posterior distribution. An efficient Gibbs sampling algorithm is developed for sampling from the posterior distribution. A bone marrow transplant data set is analyzed in details to demonstrate the proposed methodology.

Keywords Gamma frailty • Gap time • Gibbs sampler • Piecewise exponential model • Survival analysis

1 Introduction

Markov models and Markov extension models are routinely used to model multi-state data, in which the transition probabilities and transition intensities are the major study focus. There are two major time scales used for studying transition intensities: the study time since the study origin and the duration time in the current state. Based on the time scale to use, Markov model and Markov extension models are classified into several categories. The non-homogeneous Markov model assumes

M. de Castro (✉)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo,
São Carlos, SP, Brazil
e-mail: mcastro@icmc.usp.br.

M.-H. Chen
Department of Statistics, University of Connecticut, Storrs, CT, USA

Y. Zhang
Agios Pharmaceuticals, 88 Sidney St, Cambridge, MA, USA

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_4

that the upcoming transition intensity depends only on the history via the current state and the duration time. The homogeneous Markov model (classical Markov model) further assumes that all transition hazards are constant. The homogeneous semi-Markov model assumes that the transition intensity, given the current state and the duration time in the current state, is independent of the history. The non-homogeneous semi-Markov model allows the transition intensities to depend on both of the study and duration time scales. Here, we consider Markov and semi-Markov models only.

There is a rich literature on models for multi-state data. Most of published works are based on frequentist methods. Comprehensive reviews about the development and applications of multi-state models are given by, for example, Commenges (1999), Hougaard (1999), Hougaard (2000), Andersen and Keiding (2002), Andersen and Perme (2008), Meira-Machado et al. (2009), Zhao (2009), Andersen and Perme (2013), and references therein.

In the literature on analyzing bone marrow transplant (BMT) data, such as Fiocco et al. (2008) and de Castro et al. (2015), both relapse (return of leukemia) and death in remission are typically considered as terminating events, both seen as failures of the transplantation. However, relapse is not an terminating event and data are still available from relapse to death in remission. In this chapter, we extend the path-specific frailty model of de Castro et al. (2015) to the multi-state survival data with only a single terminating event. For the BMT data, we consider only death in remission as a terminating event. We then reanalyze the BMT data to investigate the consequences of omitting the transition from relapse to death in remission.

The rest of the chapter is organized as follows. Section 2 presents a description of the BMT data. The detailed development of the proposed model is given in Sect. 3. The prior distribution is specified and the propriety of the posterior distribution is established in Sect. 4. Moreover, model comparison criteria are also presented. In Sect. 5, the methodology is applied to the analysis of the bone marrow transplant data set presented in Sect. 2. We conclude the chapter with brief discussion and remarks in Sect. 6.

2 BMT Data

According to Fiocco et al. (2008), bone marrow transplantation is an effective and standard treatment for acute leukemia, but the procedure is associated with considerable morbidity and mortality. The BMT data used in this chapter are available in the mstate package in R (de Wreede et al. 2011). Six states are considered. As shown in Fig. 1, after transplant, the patients may experience platelet recovery (PR), acute graft-versus-host disease (AGvHD), both PR and AGvHD, or relapse, all of which are considered as nonterminating events. The patients eventually experience death in remission, which is of course a terminating event. Graft-versus-host disease (GvHD), either acute or chronic, is the most common non-relapse complication. Patients who develop acute GvHD are more likely to

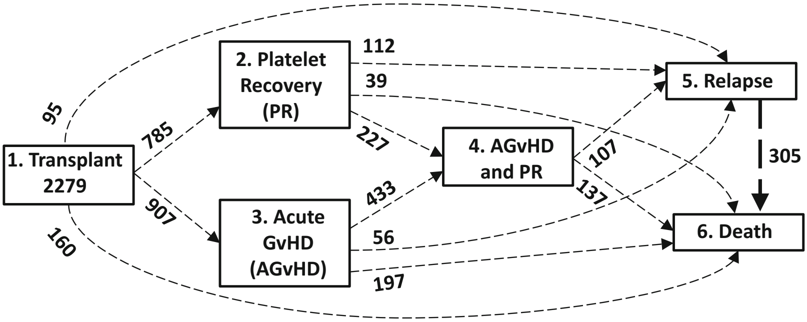


Fig. 1 Path diagram of the BMT data

develop chronic GvHD than others. Overall, there are 13 transitions and the number of patients in these transitions are shown in Fig. 1. These data motivate our proposed methodology. The data are comprised of 2279 patients treated between 1985 and 1998. Table S2 in de Castro et al. (2015) presents the baseline prognostic factors in Table 1. The reference classes for year of transplant and age at transplant (in years) are 1985–1989 and ≤20, respectively.

Unlike de Castro et al. (2015) as well as Fiocco et al. (2008), we add the transition 5 → 6 to the structure as shown in Fig. 1. We note that about 300 patients passed this transition. We will examine the consequences of omitting the transition 5 → 6.

We can also view Fig. 1 as a path diagram, with boxes representing states and arrows representing transitions from a parent state to a child state. In Fig. 1, state 5 is the only state having a single child state and there are $K = 10$ possible complete paths and their corresponding states and transitions are given as follows: P1: 1 → 5 → 6, P2: 1 → 2 → 5 → 6, P3: 1 → 2 → 6, P4: 1 → 2 → 4 → 5 → 6, P5: 1 → 2 → 4 → 6, P6: 1 → 3 → 4 → 5 → 6, P7: 1 → 3 → 4 → 6, P8: 1 → 3 → 5 → 6, P9: 1 → 3 → 6, and P10: 1 → 6. There are five possible incomplete paths as well, including IP1: 1, IP2: 1 → 2, IP3: 1 → 3, IP4: 1 → 2 → 4, and IP5: 1 → 3 → 4.

Among the 2279 patients who underwent transplant, the numbers of patients belonging to these paths are 95 patients for P1, 112 patients for P2, 39 patients for P3, 33 patients for P4, 60 patients for P5, 74 patients for P6, 77 patients for P7, 56 patients for P8, 197 patients for P9, 160 patients for P10, 332 patients for IP1, 407 patients for IP2, 221 patients for IP3, 134 patients for IP4, and 282 patients for IP5. The median gap times in years of the transitions in Fig. 1 are 0.066 for 1 → 2, 0.045 for 1 → 3, 0.627 for 1 → 5, 0.144 for 1 → 6, 0.027 for 2 → 4, 0.586 for 2 → 5, 0.479 for 2 → 6, 0.041 for 3 → 4, 0.675 for 3 → 5, 0.205 for 3 → 6, 0.578 for 4 → 5, 0.444 for 4 → 6, and 0.252 for 5 → 6.

3 Models

In this section, we introduce the two main components of our models. We begin with some definitions. If a state does not have any child states, then it is an absorbing state, whereas if a state does not have a parent state, then it is a starting state. We are interested in studies in which there is one starting state and one absorbing state. We only consider a progressive model, in which all the transitions are in one direction from an early state to a late state but not vice versa. If a state is neither a starting state nor an absorbing state, then it is a transient state. A transient state must have both parent and child states. As discussed in Sect. 2, a sequence of connected states from the starting state to the absorbing state is called a path. There are $J \geq 2$ states and $K \geq 1$ paths. We assume that each subject goes through just a single path.

3.1 Model for Immediate Child States

For each parent state j, let $\mathscr{P}_j = \{\ell : \text{state } \ell \text{ is an immediate child state of state } j\}$ denote the collection of all possible immediate child states of state j, $j = 1, \ldots, J$. We denote the starting and absorbing states as 1 and J, respectively. We also assume $\ell > j$ when $\ell \in \mathscr{P}_j$. Let δ_j denote a possible value of a child state for the parent state j, independent of the gap times. A multinomial logistic regression model is assumed for δ_j with probability

$$P(\delta_j = \ell | z_1, \boldsymbol{\alpha}^{(j)}) = \exp(z_1' \boldsymbol{\alpha}_\ell^{(j)}) / \sum_{\ell^* \in \mathscr{P}_j} \exp(z_1' \boldsymbol{\alpha}_{\ell^*}^{(j)}),$$

where z_1 is a $p \times 1$ vector of covariates, $\boldsymbol{\alpha}_\ell^{(j)}$ is a $p \times 1$ vector of parent and child specific regression coefficients for $\ell \in \mathscr{P}_j$, and $\boldsymbol{\alpha}^{(j)} = ((\boldsymbol{\alpha}_\ell^{(j)})' : \ell \in \mathscr{P}_j)'$. We assume $\boldsymbol{\alpha}_{\ell_j}^{(j)} = \mathbf{0}$, where $\ell_j = \max\{\ell : \ell \in \mathscr{P}_j\}$, to ensure identifiability.

3.2 Models for the Gap Times with Path Specific Frailty

Let $T_{j\ell}$ denote the gap time between two connected states j and ℓ. In Fig. 1, we have a total of 13 gap times. Let $S_k = \{\ell : \text{state } \ell \text{ is in path } k\}$ denote the collection of states along path Pk for $k = 1, \ldots, K$, and $|S_k|$ denote the cardinality of the set S_k. For each path Pk, in order to model variability in the hazard not accounted for observable covariates, we introduce a frailty term w, which has a path specific distribution with parameter τ_k. Each subject eventually ends up with a certain path. Each path is composed by a number of states representing intermediate points before reaching the absorbing state J. The vulnerability, represented by the frailty, depends on the sequence of events experienced by the subject.

For complete paths, we assume that

$$w|\tau_k \sim \text{gamma}(1/\tau_k, 1/\tau_k) \tag{1}$$

independently, with $E(w|\tau_k) = 1$ and $Var(w|\tau_k) = \tau_k$. The distribution of the frailty can be determined only if we know which path the transition $j \to \ell$ belongs to. We emphasize that the frailty is subject-specific and the subjects undergoing a given path have frailties that follow the same distribution.

For complete paths, it is observed that for the path $1 \to J$ the transition between the states 1 and J cannot be shared by other paths. If $|S_k| > 2$, then the transition $j \to \ell$ can be shared by other paths. In this case, there will be more than one path specific conditional distributions for the corresponding gap time $T_{j\ell}$ according to (1).

Next, given Pk, for $j, \ell \in S_k$, the path specific hazard function for $T_{j\ell}$ when states j and l are connected is assumed as $h_{j\ell}(t|z_2, \boldsymbol{\beta}_{j\ell}, \text{P}k) = wh_{j\ell 0}(t)\exp(z_2'\boldsymbol{\beta}_{j\ell})$, where the distribution of $w|\tau_k$ is given in (1), $h_{j\ell 0}(\cdot)$ is the transition specific baseline hazard function, z_2 is a $q \times 1$ vector of covariates, and $\boldsymbol{\beta}_{j\ell}$ is a vector of transition specific regression coefficients. In the case that $\delta_j \neq \ell$, we assume that $T_{j\ell} = \infty$. Notice that in the marginal model, the proportionality of the hazards is relaxed, even for the path $1 \to J$. For notational simplicity, we assume that the covariates z_1 (Sect. 3.1) and z_2 are the same for all transitions.

The baseline hazard function $h_{j\ell 0}(\cdot)$ is represented by a piecewise constant function. First we create a partition of the gap time axis with $M_{j\ell}$ intervals and cut-points $0 = c_{j\ell 0} < c_{j\ell 1} < \cdots < c_{j\ell M_{j\ell}}$, where $c_{j\ell M_{j\ell}} > t_{ij\ell}$ for all subjects i sharing the transition $j \to \ell$. In this way, the intervals are $(0, c_{j\ell 1}]$, $(c_{j\ell 1}, c_{j\ell 2}]$, $\ldots$, $(c_{j\ell M_{j\ell}-1}, c_{j\ell M_{j\ell}}]$. We also define an interval indicator $I_{ij\ell m}$ such that $I_{ij\ell m} = 1$ if a subject i sharing the transition $j \to \ell$ failed or was right-censored in the m-th interval; $I_{ij\ell m} = 0$ otherwise. In the m-th interval, we assume a constant hazard $\lambda_{j\ell m}$, $m = 1, \ldots, M_{j\ell}$, so that $h_{j\ell 0}(t) = \lambda_{j\ell m}$ and $H_{j\ell 0}(t) = \lambda_{j\ell m}(t - c_{j\ell,m-1}) + \sum_{g=1}^{m-1} \lambda_{j\ell g}(c_{j\ell g} - c_{j\ell,g-1})$, when $c_{j\ell,m-1} < t \leq c_{j\ell m}$. In the results reported in Sect. 5, for $m = 1, \ldots, M_{jl}$, we chose intervals $(c_{j\ell,m-1}, c_{j\ell m}]$ based on the percentiles of the gap times for subjects with a complete path.

3.3 Likelihood Function

Let n denote the number of subjects. For the i-th subject, let y_{ij} denote the observed event time or right-censored time at state j. When $j = 1$, which is the starting state, let $y_{i1} = 0$. Let δ_{ij} denote a possible value of child states for the parent state j. If $\delta_{ij} = \ell$, then the gap time between the parent state j and its child state $\ell \in \mathscr{P}_j$ can be expressed as $t_{ij\ell} = y_{i\ell} - y_{ij}$. This gap time can be the gap time between two events ℓ and j or the gap time between a censoring time and an event time. For a given observation i, let ν_i denote the indicator of the absorbing state, with $\nu_i = 1$ if the absorbing state is reached and $\nu_i = 0$ otherwise.

Let $\mathscr{S}_i = \{s_{i1}, s_{i2}, \dots, s_{iJ_i}\}$ denote the set of states visited by the i-th subject, comprising a complete or incomplete path with $s_{i1} = 1$, where $J_i \geq 1$ is the number of states. We let $D = (n, \boldsymbol{t}, \boldsymbol{\nu}, \mathscr{S}_1, \dots, \mathscr{S}_n, \boldsymbol{Z}_1, \boldsymbol{Z}_2)$ denote the observed data, where $\boldsymbol{t}$ is a vector with elements $t_{i,1,s_{i2}}, \dots, t_{i,s_{i,J_i-1},s_{iJ_i}}$, for $i = 1, \dots, n$, $\boldsymbol{\nu} = (\nu_1, \dots, \nu_n)'$, $\boldsymbol{Z}_1$ is the $n \times p$ matrix of covariates with the i-th row z'_{i1}, and $\boldsymbol{Z}_2$ is the $n \times q$ matrix of covariates with the i-th row z'_{i2}.

If $s_{iJ_i} = J$, $\#(\mathscr{S}_i)$ denotes the number of the path corresponding to $\mathscr{S}_i$. If $\nu_i = 0$, then s_{iJ_i} is a transient state. In this case, let $\mathscr{U}(s_{iJ_i})$ stand for the set of subpaths from state s_{iJ_i} to the absorbing state through all the possible transient states. Each subpath $g \in \mathscr{U}(s_{iJ_i})$ has states $\{s_{i,g1}, s_{i,g2}, \dots, s_{i,gJ_{ig}}\}$, with $s_{i,g1} = s_{iJ_i}$, $s_{i,gJ_{ig}} = J$, and such that $\mathscr{S}_i \cup g$ represents a complete path. For example, in Fig. 1 we see that $\mathscr{U}(3) = \{\{3,6\}, \{3,4,6\}, \{3,5,6\}, \{3,4,5,6\}\}$. For a set $\mathscr{S}_i^*$ representing a complete path ($\mathscr{S}_i^*$ can be either $\mathscr{S}_i$ or $\mathscr{S}_i \cup g$), let $a(\tau_{\#(\mathscr{S}_i^*)}) = \prod_{j=1}^{J_i-2}(1 + j\tau_{\#(\mathscr{S}_i^*)})$, if $J_i > 2$; $\quad a(\tau_{\#(\mathscr{S}_i^*)}) = 1$, if $J_i \leq 2$, noticing that $a(\cdot)$ is obtained when we integrate out w using the Laplace transformation of a gamma$(1/\tau_{\#(\mathscr{S}_i^*)} + J_i - 1, 1/\tau_{\#(\mathscr{S}_i^*)})$ random variable. If $\mathscr{S}_i$ is a complete path ($\nu_i = 1$), let $H^*(\mathscr{S}_i) = \{1 + \tau_{\#(\mathscr{S}_i)} \sum_{j=1}^{J_i-1} H_{s_{ij},s_{i,j+1},0}(t_{i,s_{ij},s_{i,j+1}}) \exp(z'_{i2}\ \boldsymbol{\beta}_{s_{ij},s_{i,j+1}})\}^{-1/\tau_{\#(\mathscr{S}_i)}-J_i+1}$. If $\mathscr{S}_i$ is an incomplete path, $\nu_i = 0$ and the gap times are right-censored at $t_{i,s_{iJ_i},*}$. In this case, let $H^*(\mathscr{S}_i \cup g) = [1 + \tau_{\#(\mathscr{S}_i \cup g)}\{\sum_{j=1}^{J_i-1} H_{s_{ij},s_{i,j+1},0}(t_{i,s_{ij},s_{i,j+1}}) \exp(z'_{i2}\boldsymbol{\beta}_{s_{ij},s_{i,j+1}})$ $+H_{s_{iJ_i},s_{i,g2},0}\ (t_{i,s_{iJ_i},*}) \exp(z'_{i2}\boldsymbol{\beta}_{s_{iJ_i},s_{i,g2}})\}]^{-1/\tau_{\#(\mathscr{S}_i \cup g)}-J_i+1}$.

The vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\lambda}', \tau^*, \eta)'$ encapsulates all the parameters in our model, where τ^* refers to the two states path $1 \rightarrow J$, if this path exists, noticing that $\tau^* = \tau_{10}$ for the data set in Fig. 1. The latent vector $\boldsymbol{\tau}$ comprising all paths with $|S| > 2$ is included in the likelihood function to ease the computations, as in de Castro et al. (2015). Using these notations and definitions, the likelihood function can be written as

$$L(\boldsymbol{\theta}|\boldsymbol{\tau}, D) = \prod_{i=1}^{n} \{a(\tau_{\#(\mathscr{S}_i)}) H^*(\mathscr{S}_i)\}^{I(\nu_i=1)} \prod_{j=1}^{J_i-1} P(\delta_{i,s_{ij}} = s_{i,j+1}|z_{i1}, \boldsymbol{\alpha}^{(s_{ij})})$$
$$\times h_{s_{ij},s_{i,j+1},0}(t_{i,s_{ij},s_{i,j+1}}) \exp(z'_{i2}\boldsymbol{\beta}_{s_{ij},s_{i,j+1}}) \Bigg\{ \sum_{g \in \mathscr{U}(s_{iJ_i})} a(\tau_{\#(\mathscr{S}_i \cup g)}) H^*(\mathscr{S}_i \cup g)$$
$$\times \prod_{j=1}^{J_{ig}-1} P(\delta_{i,s_{i,gj}} = s_{i,g(j+1)}|z_{i1}, \boldsymbol{\alpha}^{(s_{i,gj})}) \Bigg\}^{I(\nu_i=0)}, \tag{2}$$

recalling that $s_{i,gJ_{ig}} = J$. For example, for a subject i in path P5, the contribution to the likelihood function is $(1 + \tau_5)(1 + 2\tau_5)[1 + \tau_5\{H_{120}(t_{i12}) \exp(z'_{i2}\boldsymbol{\beta}_{12}) + H_{240}(t_{i24}) \exp(z'_{i2}\boldsymbol{\beta}_{24})\ + H_{460}(t_{i46})\ \exp(z'_{i2}\boldsymbol{\beta}_{46})\}]^{-1/\tau_5-3} P(\delta_{i1} = 2|z_{i1}, \boldsymbol{\alpha}^{(1)}) P(\delta_{i2} = 4|z_{i1}, \boldsymbol{\alpha}^{(2)})\ P(\delta_{i4} = 6|z_{i1}, \boldsymbol{\alpha}^{(4)})\ h_{120}(t_{i12}) h_{240}(t_{i24}) h_{460}(t_{i46}) \exp\{z'_{i2}(\boldsymbol{\beta}_{12} + \boldsymbol{\beta}_{24} + \boldsymbol{\beta}_{46})\}$.

3.4 Path Probability

Let $\mathscr{P}_j^k$ denote the subset of $\mathscr{P}_j$ such that $\mathscr{P}_j^k = \{\ell$: state ℓ is an immediate child state of state j in path Pk, noticing that $\mathscr{P}_J^k = \emptyset$. Based on the above models, the path probability can be computed as $p_k = p_k(z_1^c) = \prod_{j \in S_k, \ell \in \mathscr{P}_j^k} P(\delta_j = \ell | z_1^c, \boldsymbol{\alpha}^{(j)})$, where z_1^c denotes the fixed values of some covariates upon which we condition and $\mathscr{P}_j^k \neq \emptyset$. This is the probability that a subject having characteristics z_1^c will eventually end up with path Pk. It is worthy to mention that under our model, the path probabilities are easy to compute. These probabilities can be useful for classifying subjects with a given set of characteristics. Furthermore, these clinically important probabilities have not been examined in the literature, including, for example, Keiding et al. (2001) and Fiocco et al. (2008), for analyzing the bone marrow transplant data discussed in Sect. 2.

3.5 Relapse Free Probability

Let T_5 denote the time to reach state 5 in Fig. 1. Then, $T_5 = T_{15}$ for path P1, $T_5 = T_{12} + T_{25}$ for path P2, $T_5 = T_{12} + T_{24} + T_{45}$ for path P4, $T_5 = T_{13} + T_{34} + T_{45}$ for path P6; and $T_5 = T_{13} + T_{35}$ for path P8. Since $P(\delta_5 = 6 | z_1) \equiv 1$ in Fig. 1, the expression of the relapse free probability, $P(T_5 > t | z_1^c, z_2^c)$ with z_1^c and z_2^c as in Sect. 3.4, is the same no matter whether state 5 is a transient state or an absorbing state. Therefore, $P(T_5 > t | z_1^c, z_2^c)$ can be computed using Eq. (9) in de Castro et al. (2015). This is a nice property, which allows us to use the same computational code developed in de Castro et al. (2015) to calculate the relapse free probability with the inclusion of the transition 5 → 6 in the path diagram and facilitates the comparison of the relapse free probabilities with including and excluding the transition 5 → 6. We note that in a general model, probabilities corresponding to any event of interest can be computed in an analogous way.

4 Bayesian Inference

4.1 Prior and Posterior Distributions

To carry out Bayesian inference, we need to specify a prior distribution for $\boldsymbol{\theta}$. To this end, improper joint priors are specified for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, i.e., $\pi(\boldsymbol{\alpha}) \propto 1$ and $\pi(\boldsymbol{\beta}) \propto 1$. We assume independent gamma priors for the components of $\boldsymbol{\lambda}$ as $\lambda_{j\ell m} \sim \text{gamma}(\gamma_{j\ell m1}, \gamma_{j\ell m2})$, with $\gamma_{j\ell m1} \geq 0$, $\gamma_{j\ell m2} \geq 0$, and $\pi(\lambda_{j\ell m} | \gamma_{j\ell m1}, \gamma_{j\ell m2}) \propto \lambda_{j\ell m}^{\gamma_{j\ell m1}-1} e^{-\gamma_{j\ell m2}\lambda_{j\ell m}}$, for $m = 1, \ldots, M_{jl}$ and all transitions $j \to l$. If $\gamma_{j\ell m1} = \gamma_{j\ell m2} = 0$, we obtain a Jeffreys-type prior for $\lambda_{j\ell m}$. In order to ensure a unique marginal distribution for the gap times, we further assume that $\tau_k | \eta \sim \text{exponential}(\eta)$

independently for the paths Pk such that $|S_k| > 2$, with density function $f(\tau_k|\eta) = \exp(-\tau_k/\eta)/\eta$, for $\tau_k > 0$. Then, after integrating out the frailty w and τ_k, there will be only one marginal distribution for the time $T_{j\ell}$ even when the transition $j \to \ell$ belongs to more than one paths. For η, we take an inverse gamma distribution with $\pi(\eta) \propto (1/\eta)^{\gamma_{01}+1}\exp(-\gamma_{02}/\eta)$, where $\gamma_{01} > 0$ and $\gamma_{02} > 0$ are hyperparameters chosen so that the resulting prior is relatively noninformative. For τ^*, we assume an inverse gamma distribution with hyperparameters $\gamma_{03} > 0$ and $\gamma_{04} > 0$. Hence, the prior has the following form $\pi(\boldsymbol{\theta}) \propto \{\prod_{\forall j\to\ell}\prod_{m=1}^{M_{j\ell}}\pi(\lambda_{j\ell m})\}\pi(\tau^*)\pi(\eta)$. The corresponding posterior distribution is thus given by

$$\pi(\boldsymbol{\theta}|D) \propto \pi(\boldsymbol{\theta})\int L(\boldsymbol{\theta}|\boldsymbol{\tau},D)\prod_{k:|S_k|>2} f(\tau_k|\eta)d\tau_k.$$

Sufficient conditions for the propriety of the posterior distribution are similar to those given in de Castro et al. (2015). These conditions were satisfied for the BMT data in Sect. 2.

4.2 Bayesian Computations and Model Comparison

The analytical form of the posterior distribution in Sect. 4.1 is not available. Therefore, we develop an efficient Gibbs sampling scheme (Robert and Casella 2004) to draw samples from the posterior distribution. To this end, we introduce many latent variables and perform reparameterizations. The details of our computational development are given in de Castro et al. (2015). Bayesian computations using the Gibbs sampler were implemented in the FORTRAN language using IMSL subroutines with double precision. The convergence of the Gibbs sampler was checked using several diagnostic tools discussed in Robert and Casella (2004).

To carry out Bayesian model comparison, we consider the deviance information criterion (DIC) and the logarithm of the pseudo marginal likelihood (LPML). We define the deviance $\text{Dev}(\boldsymbol{\vartheta}) = -2\log L(\boldsymbol{\theta}|\boldsymbol{\tau},D)$, where $\boldsymbol{\vartheta} = (\boldsymbol{\theta}',\boldsymbol{\tau}')'$ and $L(\boldsymbol{\theta}|\boldsymbol{\tau},D)$ is given in (2). Let $\overline{\boldsymbol{\vartheta}}$ and $\overline{\text{Dev}} = E\{\text{Dev}(\boldsymbol{\vartheta}|D)\}$ denote the posterior means of $\boldsymbol{\vartheta}$ and $\text{Dev}(\boldsymbol{\vartheta})$, respectively. According to Spiegelhalter et al. (2002), the DIC measure is defined as $\text{DIC} = \text{Dev}(\overline{\boldsymbol{\vartheta}}) + 2p_D^{\theta}$, where $p_D^{\theta} = \overline{\text{Dev}} - \text{Dev}(\overline{\boldsymbol{\vartheta}})$ is the effective number of model parameters. The smaller the DIC value, the better the model fits the data. The posterior means $\overline{\boldsymbol{\vartheta}}$ and $\overline{\text{Dev}}$ can be estimated by $\overline{\boldsymbol{\vartheta}} = \sum_{j=1}^{B}\boldsymbol{\vartheta}_j/B$ and $\overline{\text{Dev}} = \sum_{j=1}^{B}\text{Dev}(\boldsymbol{\vartheta}_j)/B$, where $\boldsymbol{\vartheta}_1,\ldots,\boldsymbol{\vartheta}_B$ are samples from the posterior distribution. LPML is another useful Bayesian measure of goodness-of-fit, which is defined based on the conditional predictive ordinate (CPO). For the i-th observation, we define CPO as $\text{CPO}_i = \int L(\boldsymbol{\theta}|\boldsymbol{\tau},D_i)\pi(\boldsymbol{\vartheta}|D^{(-i)})d\boldsymbol{\vartheta}$, where D_i is the observed data for the i-th subject, $L(\boldsymbol{\theta}|\boldsymbol{\tau},D_i)$ is the likelihood for the i-th subject, which is the term inside the product in (2), $D^{(-i)}$ is the data with D_i deleted, and $\pi(\boldsymbol{\vartheta}|D^{(-i)})$ is the posterior density of $\boldsymbol{\vartheta}$ based on the data $D^{(-i)}$. According to

Geisser and Eddy (1979) and Gelfand and Dey (1994), an approximation is given by LPML $= \sum_{i=1}^{n} \log(\widehat{\text{CPO}}_i)$, where $\widehat{\text{CPO}}_i = [\{\sum_{j=1}^{B} 1/L(\boldsymbol{\theta}_j|\boldsymbol{\tau}, D_i)\}/B]^{-1}$. The larger the LPML value, the better the model fits the data.

5 Analysis of the BMT Data

We carry out a detailed analysis of the BMT data described in Sect. 2. The prognostic factors in Table S2 in de Castro et al. (2015) are the covariates both for the probabilities and for the hazard functions in Sects. 3.1 and 3.2. The hyperparameters for the prior distribution in Sect. 4.1 were set at $\gamma_{j\ell m1} = \gamma_{j\ell m2} = 0$, for all j, ℓ, and m, $\gamma_{01} = \gamma_{02} = 0.01$, and $\gamma_{03} = \gamma_{04} = 5$. When running the Gibbs sampling algorithm, the first 2000 iterations were discarded. Then, we performed 200,000 additional iterations with thinning equal to 20, leading to 10,000 samples for each parameter. According to the DIC and LPML values, we select the model with $M_{12} = M_{13} = M_{34} = M_{56} = 10$, $M_{15} = M_{16} = M_{24} = M_{35} = M_{36} = M_{45} = M_{46} = 4$, and $M_{25} = M_{26} = 2$ intervals as our working model. Regarding $\pi(\tau_{10})$, a sensitivity analysis with values of γ_{03} and γ_{04} in $\{1, 2, 5, 10, 50\}$ shows that the best fit is achieved when $\gamma_{03} = \gamma_{04} = 5$.

Next we present some results obtained from the samples of the posterior distribution. The posterior estimates for some parameters are displayed in Table 1. The estimates of $\boldsymbol{\beta}_{15}$ and $\boldsymbol{\lambda}_{15}$ are more precise under the one terminal event model in Fig. 1. For the model with one terminal event, the coefficient of $\boldsymbol{\beta}_{15}$ for the category year of transplant:1990–1994 has the posterior estimate not included in the 95 % highest posterior density (HPD) interval. The changes in the estimates of $\boldsymbol{\lambda}_{15}$ are worthy of attention. We see, for example, that the reduction in the posterior standard deviation ranges by a factor from 3.5 to 64.7.

With respect to $\boldsymbol{\beta}_{45}$ and $\boldsymbol{\lambda}_{45}$, the inclusion of the transition $5 \to 6$ has less impact on the posterior estimates when we compare with the estimates under the two terminal events model. The transition $1 \to 5$ belongs to only one path (P1), whereas the transition $4 \to 5$ is shared by two paths (P4 and P6), so that the former borrows more strength when transition $5 \to 6$ is included. The estimates of τ_1, τ_4, and τ_6 behave similarly. For all the remaining parameters, the differences in the estimates under these two models are negligible.

The changes in the estimates of the path probabilities are minor. For example, for path P1 and the category age at transplant: $\leq$20 years, the posterior means and 95 % HPD intervals are 0.235 (0.198, 0.270) and 0.234 (0.198, 0.269) under the models with one and two terminating events, respectively. This may be explained by the fact that the path probabilities are computed through multinomial logistic regression models and for the diagram in Fig. 1, once a patient reaches state 5, state 6 will be the next state with probability 1.

In Fig. 2 the models with one and two terminal events are compared via the relapse free probability according to donor recipient gender mismatch and age at transplant. For all levels of these two prognostic factors, the differences in the

Table 1 Posterior summaries from 10,000 replications for the two models: mean, standard deviation (SD), and 95 % highest posterior density (HPD) interval

		One terminal event			Two terminal events		
Prognostic factor	Categories	$\boldsymbol{\beta}_{15}$					
Donor recipient		Mean	SD	95 % HPD	Mean	SD	95 % HPD
gender mismatch	Yes	0.181	0.234	(−0.268, 0.645)	0.290	0.367	(−0.435, 1.017)
Prophylaxis	Yes	0.450	0.236	(−0.010, 0.915)	0.615	0.362	(−0.043, 1.401)
Year of transplant	1990–1994	0.500	0.258	(0.015, 1.011)	0.487	0.385	(−0.315, 1.214)
	1995–1998	0.353	0.318	(−0.287, 0.962)	0.337	0.449	(−0.566, 1.208)
Age at transplant	(20,40]	−0.056	0.243	(−0.522, 0.423)	−0.086	0.367	(−0.829, 0.628)
(years)	>40	−0.146	0.339	(−0.833, 0.492)	−0.017	0.517	(−1.058, 0.991)
Prognostic factor	Categories	$\boldsymbol{\beta}_{45}$					
Donor recipient		Mean	SD	95 % HPD	Mean	SD	95 % HPD
gender mismatch	Yes	0.684	0.308	(0.071, 1.277)	0.695	0.306	(0.091, 1.286)
Prophylaxis	Yes	0.121	0.305	(−0.479, 0.718)	0.126	0.309	(−0.466, 0.749)
Year of transplant	1990–1994	−1.023	0.299	(−1.574,−0.409)	−1.029	0.300	(−1.625, −0.451)
	1995–1998	−0.683	0.331	(−1.317,−0.029)	−0.681	0.330	(−1.334, −0.040)
Age at transplant	(20,40]	0.197	0.342	(−0.454, 0.879)	0.194	0.341	(−0.493, 0.839)
(years)	>40	0.847	0.397	(0.088, 1.618)	0.857	0.397	(0.113, 1.658)
	Intervals	$\boldsymbol{\lambda}_{15}$					
		Mean	SD	95 % HPD	Mean	SD	95 % HPD
	1	0.164	0.035	(0.099, 0.232)	0.174	0.043	(0.096, 0.258)
	2	0.209	0.044	(0.130, 0.301)	0.310	0.155	(0.116, 0.605)
	3	0.160	0.035	(0.098, 0.229)	0.356	0.356	(0.099, 0.958)
	4	0.014	0.003	(0.008, 0.020)	0.076	0.194	(0.008, 0.285)
	Intervals	$\boldsymbol{\lambda}_{45}$					
		Mean	SD	95 % HPD	Mean	SD	95 % HPD
	1	0.738	0.150	(0.453, 1.030)	0.743	0.149	(0.464, 1.041)
	2	1.808	0.367	(1.103, 2.533)	1.816	0.367	(1.136, 2.536)
	3	1.348	0.293	(0.813, 1.934)	1.351	0.293	(0.809, 1.918)
	4	1.167	0.363	(0.507, 1.909)	1.141	0.353	(0.468, 1.821)
	Path	$\boldsymbol{\tau}$					
		Mean	SD	95 % HPD	Mean	SD	95 % HPD
	1	0.156	0.145	($2.5\cdot10^{-5}$, 0.433)	3.917	3.545	(0.378, 11.62)
	4	0.047	0.045	($7.6\cdot10^{-6}$, 0.135)	0.044	0.043	($2.0\cdot10^{-5}$, 0.129)
	6	0.056	0.045	($1.3\cdot10^{-5}$, 0.143)	0.056	0.047	($3.2\cdot10^{-5}$, 0.150)

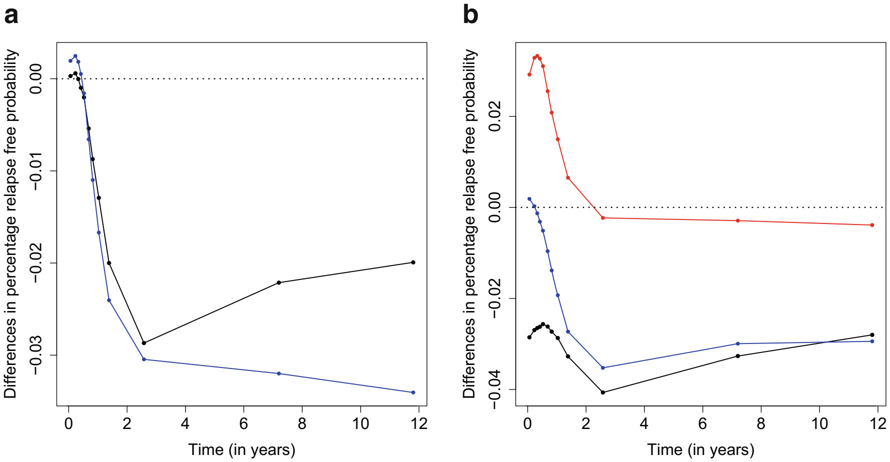


Fig. 2 Differences in percentage relapse free probability between the models with one and two terminal events. (**a**) *black*: no gender mismatch and *blue*: gender mismatch and (**b**) *black*: ≤20 years, *blue*: (20, 40] years, and *red*: >40 years

estimates are minor. Since state 5 in Fig. 1 can be reached by five paths, it would be expected that the changes in the estimates after the inclusion of the transition 5 → 6 are small. On the other side, for the four states models in de Castro et al. (2015, Sect. 6), we expect larger changes in the estimates of the probabilities.

6 Discussion

The analysis of the BMT data indicates that the omission of transition 5 → 6 potentially induces biases in the posterior estimates (see, e.g., $\boldsymbol{\beta}_{15}$ in Table 1). An extensive simulation study needs to be conducted to further confirm this. Based on our empirical investigation, linking all possible states will lead to a more valid analysis of multi-state data if data are available.

The empirical comparison of the models with one and two terminal events in Sect. 5 shows that the estimates of the relapse free probability are almost the same (see Fig. 2). However, the changes in the survival probability (state 6 in Fig. 1) are expected to be much larger since the additional five paths will be added to the calculation of the survival probability. Therefore, the omission of transition 5 → 6 may underestimate the survival probability. Even for the BMT data, the computation of the survival probability involves intractable high-dimensional integration. Thus, an efficient Monte Carlo algorithm needs to be developed for estimating the survival probability. Quantifying the magnitude of bias by omitting potentially important transitions and developing a more efficient computational algorithm of the survival probability are currently under investigation.

Acknowledgements Dr. de Castro's research was partially supported by CNPq, Brazil. Dr. Chen's research was partially supported by US NIH grants #GM 70335 and #P01 CA142538.

References

Andersen PK, Keiding N (2002) Multi-state models for event history analysis. Statistical Methods in Medical Research 11:91–115

Andersen PK, Perme MP (2008) Inference for outcome probabilities in multi-state models. Lifetime Data Analysis 14:405–431

Andersen PK, Perme MP (2013) Multistate models. In: Klein JP, van Houwelingen HC, Ibrahim JG, Scheike TH (eds) Handbook of Survival Analysis, Chapman & Hall/CRC, Boca Raton, pp 417–439

de Castro M, Chen MH, Zhang Y (2015) Bayesian path specific frailty models for multi-state survival data with applications. Biometrics 71:760–771

Commenges D (1999) Multi-state models in epidemiology. Lifetime Data Analysis 5:315–327

Fiocco M, Putter H, van Houwelingen HC (2008) Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. Statistics in Medicine 27:4340–4358

Geisser S, Eddy WF (1979) A predictive approach to model selection. Journal of the American Statistical Association 74:153–160

Gelfand AE, Dey DK (1994) Bayesian model choice: asymptotics and exact calculations. Journal of the Royal Statistical Society B 56:501–514

Hougaard P (1999) Multi-state models: a review. Lifetime Data Analysis 5:239–264

Hougaard P (2000) Analysis of Multivariate Survival Data. Springer, New York

Keiding N, Klein JP, Horowitz MM (2001) Multi-state models and outcome prediction in bone marrow transplantation. Statistics in Medicine 20:1871–1885

Meira-Machado L, de Uña Álvarez J, Cadarso-Suárez C, Andersen PK (2009) Multi-state models for the analysis of time-to-event data. Statistical Methods in Medical Research 18:195–222

Robert CP, Casella G (2004) Monte Carlo Statistical Methods, 2nd edn. Springer, New York

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society B 64:583–639

de Wreede LC, Fiocco M, Putter H (2011) mstate: an R package for the analysis of competing risks and multi-state models. Journal of Statistical Software 38:1–30

Zhao L (2009) Multi-state processes with duration-dependent transition intensities: Statistical methods and applications. Doctoral dissertation, Department of Statistics and Actuarial Science. Simon Fraser University, Canada

Bayesian Integration of In Vitro Biomarker to Analysis of In Vivo Safety Assessment

Ming-Dauh Wang and Alan Y. Chiang

Abstract Prolongation of the QT interval of electrocardiogram (ECG) is a critical safety concern in drug development. To enhance prediction of QT prolongation risks of a drug in human, it has been proposed to better integrate in vitro and in vivo models for preclinical QT prolongation assessment (Hanson et al., J Pharmacol Toxicol Methods 54:116–129, 2006). By evaluation of the Health and Environmental Sciences Institute of the International Life Sciences Institute (ILSI/HESI) data set, Chiang and Wang (Stat Biopharm Res 7:66–75, 2015) proposed a Bayesian approach to incorporation of in-vivo information in in-vitro animal QT analysis. The approach has been shown to increase predictive power, improve decision making, and reduce unnecessary exposure in animal studies. In this chapter, we extend on the previous work by Chiang and Wang (Stat Biopharm Res 7:66–75, 2015) to further investigate how in vitro data can be integrated by the proposed approach and how decisions concerning QT prolongation can be more informatively made for drugs moving from pre-clinical to clinical evaluation.

Keywords Bayesian • Posterior probability • Preclinical • Prior • QT interval

1 Introduction

Drug-induced polymorphic ventricular tachyarrhythmia, known as torsades de pointes (TdP) is a rare but potentially life threatening arrhythmia leading to syncope or even to ventricular fibrillation and sudden cardiac death, which is typically not seen in clinical trials prior to registration of a new drug (Faber et al. 1994; Belardinelli et al. 2003; Redfern et al. 2003). Occurrence of TdP has been linked to delayed cardiac repolarization, as manifested by prolongation of the QT interval on the electrocardiogram (ECG) (Algra et al. 1991). Because QT interval prolongation is frequently associated with TdP, it is considered a surrogate marker

M.-D. Wang (✉) • A.Y. Chiang
Global Statistical Sciences, Lilly Corporate Center, Eli Lilly and Company, Indianapolis, IN 46285, USA
e-mail: md_wang@lilly.com

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_5

for the potential of a drug to induce TdP (ICH E14 Gudance 2005). Since almost all drugs that produce TdP in man also inhibit the rapid form of the delayed rectifier potassium current IKr, encoded by the hERG gene, the blockade of this channel and derived electrophysiological consequences on the organ level (QT interval prolongation) are currently the primary parameters to predict drug-induced torsadogenesis. The assumption of the central role of hERG channel inhibition as the mechanism of drug-induced TdP also leads to the idea of screening new chemical entities for hERG inhibitory activity early in the drug development. The half-maximal inhibitory concentration (IC50) for hERG block is often compared to maximal plasma drug concentrations to define a preclinical cardiovascular safety margin. Earlier studies have related hERG safety margins with QTc prolongation and proarrhythmic risk based on diverse sets of preclinical and clinical study conditions and impressions (Redfern et al. 2003; De Bruin et al. 2005). The ICH S7B guidance for pharmaceutical industry (2005) describes a core battery of two preclinical assays used to predict tosadogenic potential in man. These include (1) an in vitro assay investigating the inhibitory potential of a drug on IKr, and (2) an in vivo assay in a non-rodent species (typically dog or non-human primate) evaluating changes in the QT interval of the ECG. Results of these two assays together with other relevant nonclinical information (e.g. chemical/pharmacological class) are then summarized to provide an integrated risk assessment.

To improve preclinical integrated risk assessment of TdP, Chiang and Wang (2015) proposed a Bayesian approach to incorporation of in vivo information into in vitro animal QT analysis by evaluation of the Health and Environmental Sciences Institute of the International Life Sciences Institute (ILSI/HESI) data. We further extend the work by Chiang and Wang (2015) in this paper. The original analysis employed a prior distribution of QT intervals elicited from hERG channel inhibition, and then combined it with the observed in vivo data to obtain the posterior distribution. The extension herein considered elicits the prior from the relation of change in the QT interval to exposure-adjusted rather than unadjusted hERG, which has been deemed as more relevant to prediction of QT prolongation liability (Redfern et al. 2003). Statistical inferences, including point estimates, confidence intervals and hypothesis tests, follow as appropriate summaries of the posterior (Spiegelhalter et al. 2004). While there is an established meaningful QT interval change in human (see for example, ICH E14 Gudance 2005), no consensus cut-off value has been established in preclinical setting. Another extension from Chiang and Wang (2015) is the exploration of various cut-off values of QT interval change for making preclinical decisions concerning classification of drugs for TdP causing potential in man following the Bayesian analysis.

The paper is organized as follows. The proposed Bayesian analysis model and its posterior inference are detailed in Sect. 2. In Sect. 3, we used the ILSI/HESI data set to illustrate the proposed approach. Differentiating from the previous work by Chiang and Wang (2015), prior elicitation was conducted in a modified manner. In addition, the topic of the magnitude of increase in the QT interval for classification of TdP causing potential is examined. Finally, Sect. 4 provides concluding remarks and discussion of the current extended work.

2 Methods

Most of the following details of experimental designs, data, and statistical models for this extended work can also be found in Chiang and Wang (2015).

2.1 In Vivo Telemetry Study

Twelve drugs were selected for the ILSI/HESI studies based on established association with or absence of clinical QT prolongation or TdP. Six "positive drugs" were known to cause TdP in humans: bepridil, cisapride, haloperidol, pimozide, terfenadine and thioridazine. Six "negative drugs", despite extensive use, lack the TdP association: amoxicillin, aspirin, captopril, diphenhydramine, propranolol and verapamil. Each drug was evaluated in vivo using a double Latin square design (ICH S7A Guidance 2000) where eight beagle dogs each received a vehicle control and three dose levels (low, medium, and high) of the drug on four separate dosing days.

On each dosing day, data were collected continuously for at least 30 min pre-dose and ended at 20 h post-dose. Although data were continuously acquired, 15 complexes of ECG were measured and averaged at one pre-dose and seven post-dose time intervals. Table 1 gives the corresponding time points for each study. Each of the 12 drugs was investigated using the same general study design. Careful consideration was given to assure that study laboratory personnel were blinded to the identity of the test substances with each assay having a unique blinding code for the 12 test drugs (Hanson et al. 2006).

Table 1 The seven post-dose time points of data collection in each study

Drug	Time 1 (h)	Time 2 (h)	Time 3 (h)	Time 4 (h)	Time 5 (h)	Time 6 (h)	Time 7 (h)
Amoxicillin	0.5	1	2	3	6	12	20
Aspirin	0.5	1	2	4	6	12	20
Bepridil[a]	0.5	1	2	4	6	12	20
Captopril	1	1.5	2	3.5	5	8	20
Cisapride[a]	1	2	3	4	6	12	20
Diphenhydramine	1	2	3	4	6	12	20
Haloperidol[a]	0.5	1.5	2.5	3.5	5	8	18
Pimozide[a]	1	2	3	5	8	12	20
Propranolol	1	1.5	2	4	6	12	20
Terfenadine[a]	1	1.5	2	4	6	10	20
Thioridazine[a]	1	2	3	4	6	12	20
Verapamil	0.75	1.5	2	4	6	12	20

[a] Drugs with demonstrated TdP during clinical use

2.2 Statistical Analysis Method

We focus on the analysis of QTcF interval (Fridericia correction; Fridericia 1920) to illustrate the Bayesian approach to be described; other heart rate corrected QTc intervals can be analyzed similarly. A conventional statistical model for the analysis of QTcF interval is expressed as follows:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + t_l + b_l \cdot x_{ijk} + (\alpha t)_{il} + (\beta t)_{jl} + (\gamma t)_{kl} + e_{ikl} + \varepsilon_{ijkl}, \tag{1}$$

where

y_{ijkl} is the l-th post-baseline QTcF measurement of animal j in period (day) k receiving dose i, with $i, k = 1,.., 4, j = 1,.., 8$, and $l = 1,.., 7$,
μ is the overall mean,
$\alpha_i, \beta_j, \gamma_k$ and t_l describe the main effects for dose, animal, period and time, respectively,
x_{ijk} is the baseline QTcF for animal j receiving dose i in period k,
b_l is the random slope for each time point, and
$(\alpha t)_{il}$, $(\beta t)_{jl}$ and $(\gamma t)_{kl}$ are the interactions of treatment group, animal and period with time, respectively,
$e_{ikl} \sim N\left(0, \sigma_c^2\right)$ and $\varepsilon_{ijkl} \sim N\left(0, \sigma^2\right)$,

Chiang et al. (2007). Parameter constraints that allow for a unique solution of the main effects and interactions are implicit in the model. The random errors e_{ikl}'s and ε_{ijkl}'s are assumed mutually independent to constitute a compound symmetry covariance structure for measurements on an animal over time. That is, the covariance structure for the measurements from the same animal across the time points is denoted by $\Sigma_{ikl} = \sigma^2 I_7 + \sigma_c^2 J_7$, where I_7 is the (7×7) identity matrix and J_7 is the (7×7) matrix with all entries equal to one. For this application, Chiang et al. (2007) indicated that compound symmetry covariance structures were most adequate based on the Akaike's information criterion, as compared to other covariance structures.

In drug discovery, potency of hERG block is typically used as an early screen for evolving preclinical drug candidates to avoid the risk of delayed cardiac repolarization. Redfern et al. (2003) investigated 100 drugs and suggested a negative relationship between increases in QTc in vivo and the inhibitory concentration values of hER in vitro. In the ILSI/HESI experiments, the inhibition of hERG was determined by measuring the peak amplitude of the tail currents at −80 mV before and after drug administration. At least four different concentrations of test substance were used to define the concentration–response relationships and for each of these drug concentrations three to eight different cells were examined. The IC50 was determined from a curve fit of Hill equation to the data points:

$$y = \left(100\% \times [drug]^n\right) / \left([IC_{50}]^n + [drug]^n\right),$$

where y is the percent inhibition, $[drug]$ denotes the drug serum concentration, and n is a coefficient that determines the slope of the curve. Chiang and Wang (2015) presented the scatter plot of the hERG IC50 values of the 12 test drugs at high doses conducted at ChanTest Lab and their QTcF changes from baseline. It was observed that there exists a negative relationship between the log-transformed IC50 values for hERG blockade and the QTcF changes, which suggests utilization of this prior knowledge of i correlation and available data of hERG IC50 in vitro for analysis of QTcF in vivo.

2.3 The Bayesian Model

Instead of a frequentist mixed-effects analysis (Chiang et al. 2007), Chiang and Wang (2015) proposed a Bayesian approach by incorporating knowledge of the correlation of hERG IC50 and QTcF change into the mixed-effects analysis model for QTcF. The model for its most part is re-delineated below for the sake of completeness of presenting the current extension. First, the mixed effects model (1) is re-parameterized as follows:

$$y_{ijkl} = \alpha_{il} + \beta_j + \gamma_k + t_l + b_l \cdot x_{ijk} + (\beta t)_{jl} + (\gamma t)_{kl} + e_{ikl} + \varepsilon_{ijkl}, \tag{2}$$

where α_{il} is the cell mean for treatment i and time l. The parameters are treated random and are collectively denoted by

$$\Theta = \left(\alpha_{il}\text{'s},\ \beta_j\text{'s},\ \gamma_k\text{'s},\ t_l\text{'s},\ b_j\text{'s},\ (\beta\ t)_{jl}\text{'s},\ (\gamma t)_{kl}\text{'s},\ \sigma_c^2,\ \sigma^2\right).$$

Let the prior knowledge of Θ be expressed by a density $\pi(\Theta)$. We assign prior distributions to the parameter components independently, which are assumed flat (i.e. $\propto 1$) except for α_{il} ' s, σ_c^2, and σ^2. Thus $\pi(\Theta)$ is proportional to

$$\pi\left(\alpha_{il}, \sigma_c, \sigma\right) = \left\{\prod\nolimits_{il} \pi\left(\alpha_{il}\right)\right\} \cdot \pi\left(\sigma_c^2\right) \cdot \pi\left(\sigma^2\right).$$

In Chiang and Wang (2015), prior distributions for changes in QTcF were derived from its relationship to non-exposure adjusted hERG IC50. That is, all doses of the same drug were assigned the same prior for QTcF change from baseline. However, it is exposure-adjusted hERG IC50, namely, the ratio hERG IC50 and free plasma concentration (or hERG safety margin) that is believed and has been shown to be more relevant to prediction of in vitro QT prolongation (Redfern et al. 2003). So the first adjustment to Chiang and Wang (2015) is to derive priors for QTcF change based on the linear regression of QTcF change versus log-transformed hERG safety margin than just hERG IC50. The scatter plot with its regression fit is shown in Fig. 1, which has an intercept of 6.65, a slope of −2.24, and a residual standard error of 8.80.

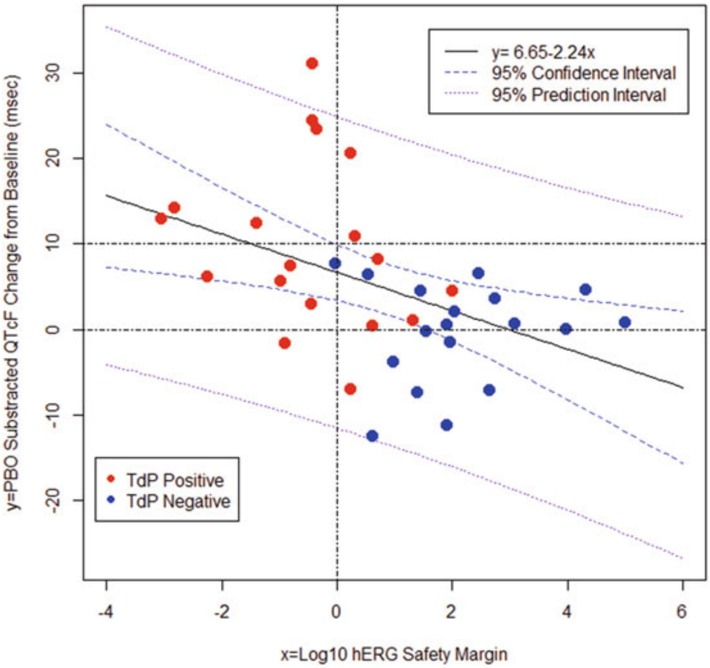


Fig. 1 Correlation of vehicle-subtracted QTcF change from baseline and hERG safety margin

Let $i = 1, 2, 3, 4$ represent vehicle, low, medium, and high dose of a test drug. Then

$$\pi\left(\alpha_{1l}\right) \propto 1; \quad \pi\left(\alpha_{il}\right) \sim N\left(\mu_{il0}, \sigma_{il0}^{2}\right),\ i = 2,\ 3,\ 4,\ l = 1, \ldots 7, \tag{3}$$

where μ_{il0} and σ_{il0}^{2} are the elicited prior mean and variance for dose i administered at time point l, respectively. The values of μ_{il0} and σ_{il0}^{2} will be specified in Sect. 3. Note that the assignment of priors is according to availability of data in the vehicle-subtracted form, which should be more variable than that of individual doses upon assumption of independence across doses.

Combining the elicited prior distributions of QTcF change with the observed in vivo QTcF data, say D, gives an updated estimation of Θ through the Bayes theory expressed by

$$\pi\left(\Theta \middle| D\right) \propto f\left(D \middle| \Theta\right) \pi\left(\Theta\right),$$

where $f(D|\Theta)$ is the likelihood of Θ given the observed QTcF data D. This derived posterior distribution enables a Bayesian estimation of drug-induced QT prolongation through inference on $\pi(\alpha_{il}\text{'}s|D)$.

2.4 Posterior Inference

According to the ICH E14 Guidance for clinical evaluation of QT/QTc, a human thorough QT (TQT) study is regarded negative if "an upper one-sided 95 % CI of QTc prolongation effect < 10 millisecond (msec)", which implies the mean must be further lower than 10 msec. For this reason a change of < 10 msec has been targeted for designing TQT studies. For example, Dong et al (2013) considered testing a mean change of $\leq$5 against >5 msec. On the earlier end, for preclinical evaluation of QT prolongation, the ICH S7B Guidance does not specify a clinically relevant threshold for increase in the QT interval. Nevertheless, in effort to improve preclinical assessment of QT prolongation, it is critical to bear in mind its relevance to clinical implications (Vargas et al. 2015; Wallis 2010). At the same time, the difficulty of the matter is recognized, as Vargas et al. (2015) states: "This is also a very challenging question to address because an increase in QT of approximately 5–10 msec constitutes a positive effect in the clinical QT study and yet there are no accepted criteria for a positive effect in the non-clinical studies". However, Vargas et al. (2015) went on to propose their evidence based suggestion that a change of 10 msec, than the commonly used ~25 msec, in animals may be more appropriate to predict 5–10 msec change in humans. With the above exemplified references, we examine the range between 5 and 10 msec of QT interval increase for decision making post the proposed Bayesian analysis.

To make inference about QT prolongation following the proposed Bayesian analysis, the posterior probability (*PP*) of QTcF change from baseline greater than a chosen "QT prolongation threshold" of $5 \leq \Delta \leq 10$ in comparison with vehicle is calculated. That is, $PP = \Pr\left(\alpha_{il} - \alpha_{1l} > \Delta \middle| D\right)$ is the vehicle-adjusted posterior probability of QTcF change from baseline exceeding Δ msec for dose $i = 2, 3, 4$ at time point l. To obtain the posterior probability for the overall time averaged QTcF increase, *PP* can be denoted as $PP = \Pr\left(\overline{\alpha}_{i.} - \overline{\alpha}_{1.} > \Delta \middle| D\right)$. If the value of *PP* is greater than a pre-specified level, say η_1, one can conclude that the drug is highly liable to cause QT prolongation in vivo. On the other hand, if the value of *PP* is less than a pre-specified level, say η_2, one may argue the drug has low risk of QT prolongation. If the value of *PP* falls between η_1 and η_2, further investigation may be needed. This information presents a key element of the Go/No-Go decision: stop further development due to preclinical safety concern, or progress to the next stage of development supported by lack of evidence in QT risk. Selection of the values of η_1 and η_2 depend on consideration of ethics and degree of risk in drug development an institution is willing to accept. For example, Dmitrienko and Wang (2006) suggested the use of $\eta_1 = 0.8$ and $\eta_2 = 0.2$ for claiming efficacy and futility respectively in clinical settings. We assimilated the suggestion to the current problem, and proposed $\eta_1 = 0.75$ and $\eta_2 = 0.25$ instead for its application to a preclinical decision. In our experience, this choice also appears to be reasonable for assessing cardiovascular safety at the preclinical stage of drug development. Otherwise, optimal selection of η_1 and η_2 driven by the ILSI/HESI data can also be statistically made by maximizing the probably of correct classification of the 12 ILSI/HESI study drugs as TdP positive or negative.

3 Analysis and Results

To specify the priors defined in (3), for the reason of lacking further detailed evidence, they are assumed independent of time and day. Then the regression fit for QTcF change versus hERG safety margin shown in Fig. 1 was utilized for the prior construction. The averaged standard deviation for the simultaneous 95 % prediction intervals for the regression line is 9.38 msec. Thus, to reflect the uncertainty of IC50 estimation, we set $\sigma_{il0} = 10$ msec so that the prior-distributions were

$$\pi\left(\alpha_{il}\right) \sim N\left(\mu_{il0}, 10^2\right), \quad i = 2,\ 3,\ 4; \quad l = 1, \ldots,\ 7,$$

where the values of the prior means μ_{il0}'s for the 12 drugs are listed in Table 2. For $i = 1$, the vehicle group, we assumed $\pi\left(\alpha_{1l}\right) \sim N\left(0, 10^2\right)$, $l = 1, \ldots,\ 7$. For the intra- and inter-animal variances, without specific information, the prior distributions were assumed to be:

$$\pi\left(\sigma_c^2\right) \sim \mathit{Gamma}\,(0.1, 0.001) \ \mathit{and}\ \pi\left(\sigma^2\right) \sim \mathit{Gamma}\,(0.1, 0.001)$$

as a variant to *Inverse Gamma* (0.001, 0.001), as described on page 170 of Spiegelhalter et al. (2004), where the mean and variance of $\mathit{Gamma}(\alpha, \beta)$ are α/β and α/β^2. Chiang and Wang (2015) showed that posterior analysis can be sensitive to prior selection in this Bayesian application, which is not further studied in the current extension.

Table 2 Prior means (vehicle-substracted) for the 12 drugs in the ILSI/HESI studies

	Prior mean (msec)		
Drug	Low dose	Medium dose	High dose
Amoxicillin	0.74	2.40	3.21
Aspirin	−4.54	−2.21	−2.87
Bepridil[a]	2.18	5.09	6.16
Captopril	2.39	4.49	5.47
Cisapride[a]	8.87	8.69	9.80
Diphenhydramine	0.52	1.19	3.43
Haloperidol[a]	6.16	7.65	8.49
Pimozide[a]	11.73	13.51	12.98
Propranolol	−0.25	2.13	3.56
Terfenadine[a]	3.74	5.31	5.98
Thioridazine[a]	7.62	7.46	7.62
Verapamil	2.28	5.28	6.76

[a] Drugs with demonstrated TdP during clinical use

Table 3 Summary of Bayesian analysis of QTcF interval using hERG IC50 elicited informative priors

	Posterior statistics	
Drug	Mean	SD
Amoxicillin	0.94	3.08
Aspirin	1.82	3.24
Bepridil[a]	11.59	4.40
Captopril	5.84	3.69
Cisapride[a]	10.88	4.12
Diphenhydramine	3.95	3.83
Haloperidol[a]	8.62	3.72
Pimozide[a]	13.45	4.11
Propranolol	3.66	5.34
Terfenadine[a]	7.86	4.20
Thioridazine[a]	15.00	4.75
Verapamil	6.53	3.63

The results are for placebo-subtracted change from baseline in msec

[a] Drugs with demonstrated TdP during clinical use

The analysis was performed by MCMC simulation using WinBUGS and R through the R BRugs package. MCMC simulations were applied here because the posterior distribution cannot be analytically derived. For details of the MCMC procedure, see Sect. 3.4 of Marin and Robert (2007). Trace and autocorrelation plots for the parameters in the MCMC were examined to ensure satisfactory conversion. Table 3 summarizes the posterior statistics, including means and standard deviations of the vehicle-adjusted QTcF change from baseline, for the average over the time points. The by-time results are not reported and further elaborated on, which by no means to exclude particular interest in certain time points or their derivatives, such as the maximum of all time points. Also, only the estimates obtained for the high dose groups are presented as they are most representative of QT liability of the examined drugs (Hanson et al. 2006) and thus of most interest in preclinical dose finding.

As preluded in Sect. 2.4, *PP* was calculated for "QT prolongation thresholds" of 5, 6, 7, 9, 10 msec, and the values for the 12 drugs are displayed in Fig. 2. Already explained in Sect. 2.4, we used "decision thresholds" of $\eta_1 = 0.75$ and $\eta_2 = 0.25$ to assess QT prolongation risk. A first observation from the figure is that none of the drugs is classified as TdP positive by one "QT prolongation threshold" and simultaneously as negative by another. The matter is then what Δ's would leave fewer drugs in the inconclusive zone between $\eta_2 = 0.25$ and $\eta_1 = 0.75$ and also keep both sensitivity and specificity in check, as indicating higher classification power. It is seen that $\Delta = 9$ and 10 tend to over drive for higher sensitivity at the loss of specificity, whereas $\Delta = 5$ and 6 incline for higher specificity on sacrifice of sensitivity. Be attentive to both ends, $\Delta = 7$ and 8 seem better choices, which are

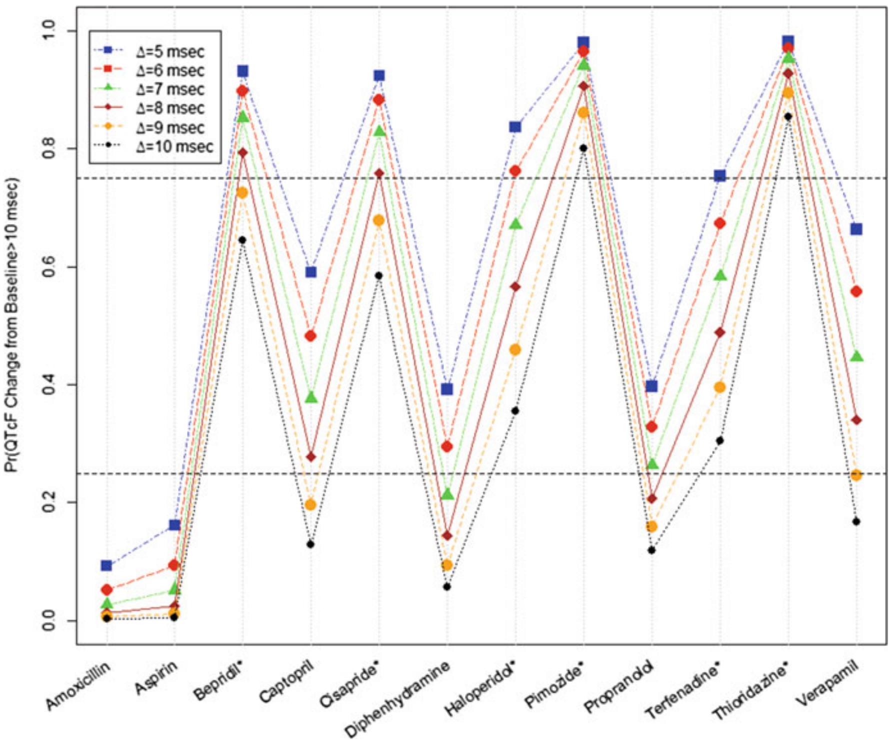


Fig. 2 Posterior probabilities of vehicle-subtracted QTcF change from baseline greater than $\Delta = 5, 6, 7, 8, 9, 10$ msec at high doses

also congruent to a mean value that would be assumed to power a TQT study to exhibit "an upper one-sided 95 % CI of QTc prolongation effect < 10 msec" (ICH E14 Gudance 2005).

One may choose to narrow the window between $\eta_1 = 0.75$ and $\eta_2 = 0.25$ to allow lower probability of falling in the inconclusive zone, but this needs to be done in consideration of inducing increased chance of misclassification. Moreover, sensitivity, specificity, and probability of inconclusive classification may bear different weights in the selection of Δ. Note that we analyzed each of 12 ILSI/HESI drugs using the same prior information summarized from all of the 12 drugs. To avoid double use of information, the analyzed drug could be left out from the construction of the correlation of QTcF change and log-transformed hERG safety margin for prior elicitation. Further exploration shows that this would not alter the results or points presented.

4 Discussion

We extended the approach proposed by Chiang and Wang (2015) on Bayesian repeated measures procedure to refine the prior elicitation and evaluate optimal dichotomization of QT interval changes in dogs for decisions concerning QT prolongation. In place of non-exposure adjusted hERG IC50, free plasma concentration adjusted hERG IC50, was explored for its correlation with in vivo QT interval change from the same drugs in the ILSI/HESI dataset. The inferred relationship was then utilized to derive priors for the proposed Bayesian in vivo QT analysis. This adjustment reflected reported literature data on the relevance of in vitro hERG inhibitory activity to in vivo QT prolongation.

The Bayesian analysis entails a posterior probabilistic quantification of QT increase greater than a meaningful threshold. While an increase of 10 msec has been often adopted as a recognized cut-off value in both clinical and preclinical settings (Chiang and Wang 2015), we examined other cutoff values than 10 msec and compared their impacts on the human TdP classification power in terms of sensitivity, specificity, and odds of inclusiveness. By the proposed Bayesian analysis of the ILSI/HESI dataset, we conclude that a cutoff value of 7−8 msec may be more appropriate for dogs as supported by the higher human TdP classification power as compared in Fig. 2. However, we must also recognize that the ILSI/HESI dataset consists of only 12 drugs, which could be questioned for generalization of our findings. Further validation can be made upon availability of more extensive data.

It has been well recognized that in vivo QT prolongation of a drug is correlated with its in vitro hERG inhibitory activity, which can be used to predict TdP liability in the clinical setting in lack of clinical QT data. If this relationship is appropriately ascertained, it can provide to enhance the design, analysis, and interpretation of an in vivo QT study by the proposed Bayesian approach (Chiang and Wang 2015). When it comes to making preclinical decisions on drug-induced TdP causing potential, the Bayesian paradigm presents a more intuitive procedure by means of posterior distributions and probabilities for the likelihood of QT prolongation than a conventional frequentist framework.

Acknowledgements The authors thank Karen Price of Eli Lily and Company and the two referees for their reviews and helpful comments, whereby the paper was improved to its current form.

References

Algra, A., Tijssen, J. G., Roelandt, J. R., Pool, J., and Lubsen, J. (1991). QTc prolongation measured by standard 12-lead electrocardiography is an independent risk factor for sudden death due to cardiac arrest, *Circulation, 83*, 1888–1894.

Belardinelli, L., Antzelevitch, C., and Vos, M.A. (2003). Assessing predictors of drug-induced torsade de pointes, *TRENDS in Pharmacological Sciences*, *24*, 619–625

Chiang, A.Y., Bass, A.S., Cooper, M.M., Engwall, M.J., Menton, R.G. and Thomas, K. (2007). ILSI-HESI Cardiovascular safety subcommittee dataset: An analysis of the statistical properties of QT interval and rate corrected QT interval (QTc), *Journal of Pharmacological and Toxicological Methods*, 56, 95–102.

Chiang, A.Y., Wang, M.-D. (2015). Combining in vitro and in vivo Cardiovascular Safety Biomarkers for Preclinical Integrated Risk Assessment. *Statistics in Biopharmaceutical Research* 7: 66–75.

De Bruin, M.L., Pettersson, L.M., Meyboom, R.H.B., Hoes, A.W., and Leufkens, H.G.M. (2005). Anti-HERG activity and the risk of drug-induced arrhythmias and sudden death. *European Heart Journal*, 26, 590–597.

Dmitrienko, A. and Wang, M.-D. (2006). Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in Medicine*, 25, 2178–2195.

Dong, X., Ding, X., Tsong, Yi. (2013). Bayesian approach to assay sensitivity analysis of thorough QT trials. *Journal of Biopharmaceutical Statistics*, 23, 73–81.

Faber, T.S., Zehender, M., and Just, H. (1994). Drug-induced torsade de pointes: Incidence, management and prevention, *Drug Safety, 11*, 463–476.

Fridericia, L.S. (1920). Die systolendauer im elektrokardiogramm bei normalen menchen und bei herzkranken, *Acta Medica Scandinavica, 53*, 469–486.

Hanson, L.A., Bass, A.S., Gintant, G., Mittelstadt, S., Rampe, D., and Thomas, K. (2006). ILSI-HESI cardiovascular safety subcommittee initiative: Evaluation of three non-clinical models of QT prolongation, *Journal of Pharmacological and Toxicological Methods*, *54*, 116–129.

ICH E14 Gudance. (2005). *Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs.*

ICH S7A Guidance. (2000). *Safety Pharmacology Studies for Human Pharmaceuticals.*

ICH S7B Guidance. (2005). *Guideline on Safety Pharmacology Studies for Assessing the Potential for Delayed Ventricular Repolarization (QT Interval Prolongation) by Human Pharmaceuticals.*

Marin, J.M., and Robert, C. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York, NY: Springer-Verlag.

Redfern, W.S., Carlsson, L., Davis, A.S., Lynch, W.G., MacKenzie, I., Palethorpe, S., Siegl, P.K.S., Strang, I., Sullivan, A.T., Wallis, R., Camm, A.J., and Hammond, T.G. (2003). Relationships between preclinical cardiac electrophysiology, clinical QT interval prolongation and torsade de pointes for a broad range of drugs: evidence for a provisional safety margin in drug development, *Cardiovascular Research,* 32–45.

Spiegelhalter D.J., Abrams K.R., and Myles, J.P. (2004). *Bayesian Approaches to Clinical Tials and Health-Care Evaluation*. West Sussex, UK: John Wiley & Sons.

Vargas, H.M., Bass, A.S., Koerner, J., Matis-Mitchell, S., Pugsley, M.K., Skinner, M., Burnham, M., Bridgland-Taylor, M., Pettit, S., Valentin, J.P. (2015). British Journal of Pharmacology, 172, 4002–4011.

Wallis, R.M. (2010). Integrated risk assessment and predictive value to humans of non-clinical repolarization assays. British Journal of Pharmacology, 159, 115–121.

A Phase II Trial Design with Bayesian Adaptive Covariate-Adjusted Randomization

Jianchang Lin, Li-An Lin, and Serap Sankoh

Abstract Adaptive randomization (e.g. response-adaptive (RA) randomization) has become popular in phase II clinical research because of its flexibility and efficiency, which also have the patient centric advantage of assigning fewer patients to inferior treatment arms. However, these designs lack a mechanism to actively control the imbalance of prognostic factors, i.e. covariates that substantially affect the study outcome. Improving the balance of patient characteristics among the treatment arms could potentially increases the statistical power of the trial. We propose a phase II clinical trial design that is response-adaptive and that also actively balances the covariates across treatment arms. We then incorporate this method into a sequential RA randomization design such that the resulting design skews the allocation probability to the better treatment arm, and also controls the imbalance of the prognostic factors across the arms. The proposed method extends the existing randomization procedures which either requires polytomizing continuous covariates or uses fixed allocation probability to adjust covariates imbalance. Simulation studies are also conducted to examine the operating characteristics of the design with existing approaches to illustrate the recommendation for clinical practice.

Keywords Adaptive randomization • Clinical trials • Bayesian adaptive design

1 Introduction

Randomization, the random assignment of clinical trial participants to different treatment arms, ensures that the observed treatment effect is attributable to the treatment itself rather than to confounding elements. An allocation procedure, randomizing entering patients based on the accrued data so far, is referred to as adaptive. According to FDA draft guidance, 2010, adaptive randomization is

J. Lin (✉) • S. Sankoh
Takeda Pharmaceuticals, 300 Massachusetts Avenue, Cambridge, MA 02139, USA
e-mail: Jianchang.Lin@takeda.com

L.-A. Lin
Merck Research Laboratories, 126 East Lincoln Avenue, Rahway, NJ 07065, USA

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_6

a form of treatment allocation in which the probability of patient assignment to any particular treatment is adjusted based on repeated comparative analysis of the allocation and response data accrued so far. Naturally, the randomization schedule across the study can change frequently or continuously over the duration of the study. However, such randomization is adopted when the outcomes are immediate and are observed faster than the study enrolment.

1.1 Response-Adaptive Randomization

The response-adaptive (RA) randomization scheme has become popular in clinical research because of its flexibility and efficiency. Based on the accruing history of patients' responses to treatment, the RA randomization scheme adjusts the future allocation probabilities, thereby allowing more patients to be assigned to the superior treatment as the trial progresses. As a result, RA randomization can offer significant ethical and cost advantages over equal randomization.

1.2 Covariate-Adaptive Randomization

Response-adaptive randomization designs have the advantage of assigning fewer patients to inferior treatment arms. However, these designs lack a mechanism to actively control the imbalance of prognostic factors, i.e. covariates that substantially affect the study outcome, across treatment arms. To ensure that any observed treatment effect is attributable to the treatment itself rather than to any particular patient characteristic, the research design must balance the potentially confounding patient characteristics among the different treatment arms. Improving the balance of patient characteristics among the treatment arms also potentially increases the statistical power of the trial. This may not be a serious issue under large samples since asymptotically the randomization automatically balances prognostic factors among treatment groups. However, for trials with small or moderate sample sizes, the imbalance of the prognostic factors can be substantial when using RA randomization designs, and thus causes difficulties to the inference after randomization. For example, in the presence of imbalanced prognostic factors, a direct comparison of marginal efficacy among the treatment arms is biased. Figure 1 illustrates that covariates can exhibit large differences between treatments as the sample size in the trial decreases.

Without considering response, various methods have been proposed to balance covariate distributions across treatment arms during randomization. For a small set of discrete covariates, stratified randomization is an effective method to achieve balance with respect to the covariates across treatment arms. This method, however, breaks down when there is a large number of covariates. Covariate-adaptive randomization (CA) designs have been developed to address this issue. In particular,

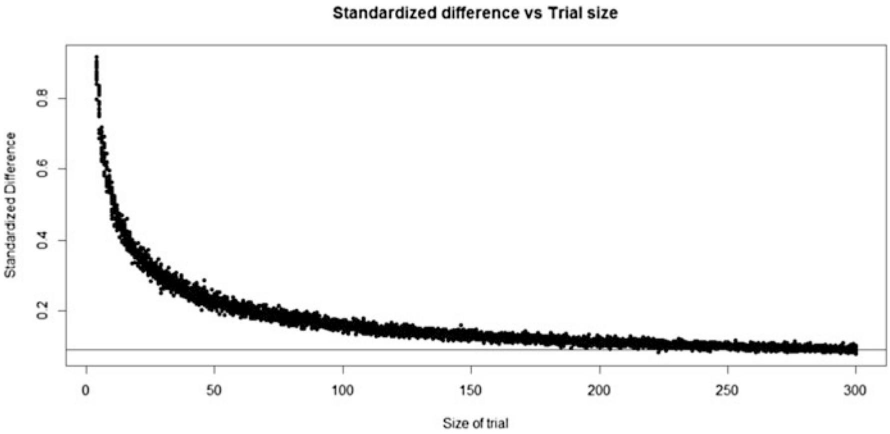


Fig. 1 Standardized difference of a N(0,1) covariate between treatments by trial size, where standardized difference = |mean(treatment)−mean(control)|/(pooled standard deviation)

Pocock and Simon (1975) proposed a minimization design to balance prognostic factors in randomization. Wei (1978) discussed the use of an urn model for CA randomization. Atkinson (1982) proposed optimal biased-coin designs for clinical trials by employing the D-optimality criterion with a linear model. Signorini et al. (1993) and Heritier et al. (2005) proposed CA randomization procedures that balance interactions between factors when such interactions exist. Scott et al. (2002), McEntegart (2003) and Lin et al. (2016) provided comprehensive reviews on CA randomization.

1.3 Proposed Covariate-Adjusted Randomization

We propose a randomization procedure that is response-adaptive and that also actively balances the covariates across treatment arms. Specifically, we develop a new covariate adaptive randomization method which assign more patients to treat arm that minimize the probability of covariate imbalance. We then incorporate this method into a sequential RA randomization design such that the resulting design skews the allocation probability to the better treatment arm, and also controls the imbalance of the prognostic factors across the arms. The proposed method extends the existing randomization where Ning and Huang (2010)'s approach requires polytomizing continuous covariates and Yuan et al. (2011)'s approach uses fixed allocation probability to adjust covariates imbalance.

2 Method

2.1 Response-Adaptive Treatment Allocation Probability

Patients are enrolled in sequential groups of size $\{N_j\}$, $j = 1, \ldots, J$, where N_j is the sample size of the sequential group j. Typically, before conducting the trial, researchers have little prior information regarding the superiority of the treatment arms. Therefore, initially, for the first j' groups, e.g. $j' = 1$, patients are allocated to K treatment arms with an equal probability $1/K$. The response information observed from these patients then can be used to skew the allocation probability in subsequent groups. Let p_k be the response rate of treatment k, and assign p_k a prior distribution of beta (α_k, β_k), for $k = 1, \ldots, K$. If, among n_k subjects treated in arm k, we observe y_k responses, then

$$Y_k \sim \text{binomial}\,(n_k, p_k) \tag{1}$$

and the posterior distribution of p_k is

$$p_k \Big|\ \text{data} \sim \text{beta}\,(\alpha_k + x_k,\ \beta_k + n_k – x_k) \tag{2}$$

During the trial, we continuously update the posterior distribution of p_k, and allocate the next patient to the kth treatment arm according to the posterior probability that treatment k is superior to all others

$$P_{RA}\,(k) = \Pr\left(p_k = \max\{p_l,\ 1 \le l \le K\}\,\Big|\text{data}\right) \tag{3}$$

2.2 A Measure of Covariate Imbalance

The measure of the degree of covariate imbalance should able to: (1) Applicable for both categorical and continuous covariates. The current method (Ning and Huang 2010) can only be used for categorical covariates. Continuous covariates need to be categorized, and it is not always clear how many categories and what cutoff values should be used. (2) Prioritize covariates that need to balance. Some prognostic factors are considered more important than others; it is desirable to assign larger weights to the more important factors when determining the overall imbalance during a randomization procedure. Yuan (2011) proposed a prognostic score measure that can accommodate both requirements.

Let $\boldsymbol{x}$ denote a vector of covariate that can be continuous or categorical, y denote the binary outcome variable, and z denote the treatment arm indicator. They assume a standard logistic regression model

$$\mathrm{logit}\left(\Pr\left(y = 1 \middle| x, z\right)\right) = \alpha + \mathbf{x}\boldsymbol{\beta}' + \gamma z \tag{4}$$

Where α, β and γ are unknown parameters. And, the prognostic score is defined as

$$\omega\left(\mathbf{x}\right) = \mathbf{x}\boldsymbol{\beta}' + \gamma z \tag{5}$$

The prognostic score automatically accommodates continuous and categorical prognostic factors, and assigns weights to prognostic factors according to their importance in predicting the response. Therefore, to balance out the effect of prognostic factors across treatment arms, we actually only need to balance the distribution of the prognostic score during the randomization.

During randomization, we assign an incoming patient to the treatment arm such that the imbalance of the prognostic score across the treatment arms is minimized. To achieve this objective, we use the Kolmogorov-Simirnov (KS) statistic as a measure of imbalance between two treatment arms. Let $\boldsymbol{w}_k$ denote the vector of prognostic scores for patients assigned to the kth treatment arm, and $S_{kk'}$ denote the KS statistic based on $\boldsymbol{w}_k$ and $\boldsymbol{w}_{k'}$ for $k \neq k'$. Then the overall imbalance among K treatment arms is measured by

$$S = \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} S_{kk'}, \tag{6}$$

2.3 Covariate-Adjusted Treatment Allocation Probability

Let $S^{(k)}$ denote the value of S if the incoming new patient is assigned to the kth treatment arm where smaller value of $S^{(k)}$ indicate less imbalance. Thus, the value of $S^{(k)}$ can be used to calculate the posterior probability that assigning this patient to treatment k minimized the overall covariate imbalance

$$P_{CA}(k) = \Pr\left(S^{(k)} = \min\left\{S^{(l)},\ 1 \le l \le K\right\} \middle| \mathrm{data}\right) \tag{7}$$

Without the prior information, the non-informative prior can be used to obtain the posterior distribution. In clinical practice, the covariates that needs to balance, are often known and pre-specified before the trial. And, the history data on the covariate effect are often available. Therefore, such prior information can be used to determine which prognostic factors need to be balanced and what's the effect size in the model (4) before conducting the randomization. Under the Bayesian framework, we elicit an informative prior of β based on historical data, and continuously update the posterior mean of β using the observed data during the ongoing trial.

2.4 Response-Adaptive Covariate-Adjusted (RACA) Treatment Allocation Probability

The idea of RACA randomization is to allow new incoming patients a better chance of being allocated to a superior treatment regimen based on cumulative information from previous patients, and adjust the allocation according to individual covariate information. Specifically, we assign a new patient to treatment k with probability $\mathrm{P}_{\mathrm{RACA}}(\mathrm{k})$,

$$\mathrm{P}_{\mathrm{RACA}}\,(\mathrm{k}) = \frac{\mathrm{P}_{\mathrm{RA}}^{\tau 1}\,(\mathrm{k})\,\mathrm{P}_{\mathrm{CA}}^{\tau 2}\,(\mathrm{k})}{\sum_{l=1}^{\mathrm{K}} \mathrm{P}_{\mathrm{RA}}^{\tau 1}(l)\mathrm{P}_{\mathrm{CA}}^{\tau 2}(l)} \tag{8}$$

Where τ1 and τ2 are the tuning parameters. There are two purposes for using tuning parameters. Firstly, we use the tuning parameter τ1 and τ2 to control the AR rate; if $\tau 1 = \tau 2 = 0$, then $\mathrm{P}_{\mathrm{RACA}}\,(\mathrm{k}) = 1/\mathrm{K}$, leading to ER. A larger value of tuning parameters would lead to a higher imbalance in allocation of patients between the arms and vice versa. Secondly, we can set different values of τ1 and τ2 to control the preference of RA and CA. If $\tau 1 < \tau 2$, we assign more weight to CA than RA, and vice versa. Furthermore, RACA randomization equivalent to RA if $\tau 2 = 0$, and CA if $\tau 1 = 0$.

If both the covariate imbalance and ethical criteria favor the assignment of a patient to the same treatment, then the new patient will be assigned to that treatment with a higher probability compared with the probability when using the simple RA or CA randomization schemes. Otherwise, the randomization procedure will result in an assignment probability between P_{RA} and P_{CA}

2.5 Early Stopping and Decision Rules

- Futility: if Pr ($\mathrm{p}_k < p.min$ | data) $> \theta_u$, where p.min denotes the clinical minimum response rate, that is, there is strong evidence that treatment k is inferior to the clinical minimum response rate, we drop treatment arm k.
- Superiority: if Pr ($\mathrm{p}_k > p.target$ | data) $> \theta_l$, where p.target denotes the target response rate, that is, there is strong evidence that treatment k is superior to prespecified response rate, we terminate the trial early and claim the treatment *k is promise*.
- At the end of the trial, if Pr ($\mathrm{p}_k > p.min$ | data) $> \theta_t$, then treatment k is selected as the superior treatment. Otherwise, the trial is inconclusive.

To achieve desirable operating characteristics, we use simulations to calibrate the pre-specified cut-off points θ_u , θ_l , and θ_t.

3 Simulation Study

We conducted simulations to evaluate the performance, under various clinical scenarios, of the proposed RACA design (two setting: RACA2 with $\tau 1 = \tau 2 = 1$, RACA3 with $\tau 1 = 1$ and $\tau 2 = 2$) to compare it with the following designs: simple equal randomization (ER), CA randomization, RA randomization, and RACA design with Yuan's method (RACA1). The patient assignment probabilities under the CA design is determined by (7) without consideration on previous patient's responses. That under the RA design is specified by (3) without considering covariate distributions. We used sample sizes of 90 for each scenario. We assigned the first 15 patients equally to three treatments (A, B, or C) and started using the adaptive randomization at the 16th patient.

We generated data from the following model,

$$\begin{aligned}\mathrm{logit}\,(\Pr(y = 1|x, z)) = \alpha + \beta 1 \text{ gender} + \beta 2 \text{ age} + \beta 3 \text{ bmi} \\ + \beta 4 \text{ race} + \beta 5 \text{ trt1} + \beta 6 \text{ trt2}\end{aligned}$$

where trt1 and trt2 is treatment indicator variable with $\text{trt1} = 1$ for treatment B, $\text{trt2} = 1$ for treatment C, and $\text{trt1} = \text{trt2} = 0$ for treatment A. We generated the continuous variable of age from uniform distribution with $\min = 25$, $\max = 80$, and BMI from N(35, $\text{sd} = 10$). Two binary indicator variable, gender and race, are generated from Bernoulli distributions with success probabilities of 0.6 and 0.3. The values of $\alpha, \beta_1, \beta_2, \beta_3, \beta_4$ are set to be –0.28, 2.1, –0.144, 0.08, and –1.5, respectively. In scenario 1, by setting $\beta_5 = \beta_6 = 0$ we obtain no differential treatment with clinical minimal response rate ($p_1 = p_2 = p_3 = 0.1$. Then we contrasted our proposed design with other randomization designs when treatment B and C was the superior treatment with a higher response rate. We set $\beta_5 = 1.378$ d $\beta_6 = 2.428$ in scenario 2, corresponding to ($p_1 = 0.1, p_2 = 0.2, p = 0.3$). The minimum clinical response rate (*p.min*) is 0.1 and the target response rate (*p.target*) is 0.25. At the end of the study, the null hypothesis of equal treatment efficacy is rejected if Pr ($p_k > p.min$ | data) > *0.9*. A total of 1,000 independent simulations were performed for each setting and randomization method. Note that when stopping rules are applied, the actually used sample size varies under different designs, which makes the comparison between designs difficult. To facilitate the comparison, we carried out simulations both with and without early stopping.

Table 1 shows the simulation results without early stopping based on the fixed sample size of 90. For each design, we list the average number of patients (with standard deviation) assigned to each treatment arm, the chance of a treatment being selected as promising, the average number of patients who achieved treatment success, the average degree of imbalance in terms of prognostic score, and the percentage of significant imbalance (the p-value of KS statistics less than 0.05).

In scenario1 (response rate equal to p.min), all designs assigned an equal number of patients to each treatment arm. However, the variations in the number of patients assigned were quite different where ER design has the smallest variation among all

Table 1 Simulation results without early stopping

Method	Arm	Response rate	# of patient assigned (SD)	Pr(selected)	# of positive response	Degree of imbalance	Percentage of significantly imbalanced covariate
Scenario1: p1 = p2 = p3 = 0.1							
ER (τ1 = τ2 = 0)	A	0.1	30.21 (4.17)	0.077	8.91	0.26	0.128
	B	0.1	29.97 (4.04)	0.077			
	C	0.1	29.83 (4.21)	0.075			
RA (τ1 = 1, τ2 = 0)	A	0.1	29.17 (17.51)	0.053	8.92	0.25	0.127
	B	0.1	30.70 (18.01)	0.054			
	C	0.1	30.13 (17.97)	0.061			
CA (τ1 = 0, τ2 = 1)	A	0.1	30.11 (4.63)	0.061	8.93	0.15	0.002
	B	0.1	30.12 (4.44)	0.061			
	C	0.1	29.77 (4.45)	0.062			
RACA1	A	0.1	30.70 (14.61)	0.068	8.95	0.19	0.023
	B	0.1	29.36 (14.08)	0.062			
	C	0.1	29.94 (14.24)	0.067			
RACA2 (τ1 = τ2 = 1)	A	0.1	29.70 (11.71)	0.056	9.15	0.17	0.002
	B	0.1	30.35 (11.45)	0.061			
	C	0.1	29.95 (11.88)	0.06			
RACA3 (τ1 = 1, τ2 = 2)	A	0.1	30.54 (9.05)	0.053	9.1	0.16	0.002
	B	0.1	29.38 (8.82)	0.053			
	C	0.1	30.08 (9.08)	0.058			

Scenario2: $p1 = 0.1$, $p2 = 0.2$, $p3 = 0.3$							
ER ($\tau 1 = \tau 2 = 0$)	A	0.1	30.11 (4.15)	0.058	17.54	0.25	0.124
	B	0.2	30.00 (4.04)	0.587			
	C	0.3	29.89 (4.23)	0.969			
RA ($\tau 1 = 1$, $\tau 2 = 0$)	A	0.1	15.75 (7.84)	0.048	21.28	0.23	0.114
	B	0.2	26.55 (15.98)	0.422			
	C	0.3	47.69 (17.18)	0.856			
CA ($\tau 1 = 0$, $\tau 2 = 1$)	A	0.1	29.84 (4.37)	0.061	17.96	0.14	0
	B	0.2	30.35 (4.47)	0.59			
	C	0.3	29.81 (4.57)	0.949			
RACA1	A	0.1	16.90 (6.88)	0.04	21.08	0.18	0.02
	B	0.2	26.26 (11.86)	0.486			
	C	0.3	46.84 (13.17)	0.942			
RACA2 ($\tau 1 = \tau 2 = 1$)	A	0.1	18.11 (5.68)	0.047	20.87	0.16	0
	B	0.2	27.45 (9.01)	0.499			
	C	0.3	44.44 (10.28)	0.958			
RACA3 ($\tau 1 = 1$, $\tau 2 = 2$)	A	0.1	20.61 (5.82)	0.049	20.66	0.15	0
	B	0.2	27.54 (8.19)	0.53			
	C	0.3	41.85 (9.48)	0.964			

designs, variation of CA design is very close to ER design, RA design has the largest variation, and the magnitude of three RACA design's variation is between CA and RA design. Without control the covariate among treatment, there are 12.8 % chance that the covariate between treatments will be significantly different at the end of trial for ER design, and 12.7 % for RA design. All three RACA design can achieve low percentage of significant covariate imbalance where the proposed method perform better than Yuan's method and we can obtain even smaller percentage of significant covariate imbalance via adjusting the values of τ1, τ2.

In scenario II, the response rates are different across treatment ($p_1 = 0.1, p_2 = 0.2$ and $p_3 = 0.3$). In contrast to the ER and CA designs, the RA and all three RACA design assigned less patients to the inferior treatment A. Comparing the patients' positive response, ER and CA designs achieved smallest number of positive response, and RA designs got highest number. The number of positive response was slightly reduced by adding CA to RA design. However, all three RACA designs got the degree of covariates imbalance reduced and achieved higher statistical power than RA design. For example, for treatment C, the power of RA design was 0.856, while that of the RACA2 design was 0.958. And, the proposed method obtained smaller degree of covariate imbalance than Yuan's method while preserved larger statistical power. It is worth to note that changing τ2 from 1 to 2 achieved less degree of covariate imbalance while assigning more patients to inferior treatment. The choice of τ1 and τ2 depends on the trial setting and the consideration of ethical and statistical issues.

Table 2 shows simulation results with early stopping ($\theta_u = \theta_l = 0.9$). In the presence of early stopping, the actual sample sizes used in trials vary under different designs. Therefore, in addition to the summary statistics that are similar to those listed in Table 2, we also reported the average sample size across 1000 simulated trials. The simulation results are similar to those achieved without stopping rules. Compared with the RA design, the proposed RACA design has a substantially lower percentage of significantly imbalanced covariates and higher statistical power. For example, in scenario II, the power under the RA design were 0.439 for treatment B and 0.818 for treatment C, while that under the RACA (τ1 = τ2 = 1) design was 0.525 and 0.878 respectively. Moreover, compared with the CA designs, the RACA design a*l*located fewer patients to the inferior treatment. For example, in scenario II, the number of patients assigned to the inferior treatment was 20.28 under the CA designs, while that under the proposed RACA (τ1 = τ2 = 1) design was only 17.94.

In summary, the simulation results show that the proposed RACA design successfully combined the advantages of the RA and the CA designs. Like the RA design, RACA design effectively skewed the allocation probability toward the superior arm. It allocated substantially fewer patients to the inferior treatment arm compared with the ER and CA designs. On the other hand, in terms of balancing the covariates, the performance of RACA design was comparable to the CA designs. A better balance of covariates under the RACA design often translated into lower type I error rate (when the efficacy of treatments are the same) or a higher statistical power (when the efficacy of treatments are different). Moreover, the proposed RACA design performs better than Yuan's method in term of covariate balancing

Table 2 Simulation results with early stopping

Method	Arm	Response rate	Sample size	# of patient assigned (SD)	Pr(selected)	Degree of imbalance	Percentage of significantly imbalanced covariate
Scenario1: $p1 = p2 = p3 = 0.1$							
ER ($\tau 1 = \tau 2 = 0$)	A	0.1	60.12	19.85 (13.69)	0.056	0.35	0.11
	B	0.1		20.11 (13.71)	0.061		
	C	0.1		20.15 (13.69)	0.057		
RA ($\tau 1 = 1$, $\tau 2 = 0$)	A	0.1	58.06	19.30 (14.18)	0.061	0.33	0.097
	B	0.1		19.67 (14.19)	0.063		
	C	0.1		19.09 (13.98)	0.06		
CA ($\tau 1 = 0$, $\tau 2 = 1$)	A	0.1	61.01	20.89 (13.69)	0.06	0.29	0.018
	B	0.1		20.11 (13.58)	0.061		
	C	0.1		20.59 (13.65)	0.063		
RACA1	A	0.1	57.01	18.45 (14.13)	0.064	0.34	0.079
	B	0.1		19.38 (14.03)	0.068		
	C	0.1		19.16 (14.17)	0.062		
RACA2 ($\tau 1 = \tau 2 = 1$)	A	0.1	57.39	19.00 (13.95)	0.051	0.32	0.038
	B	0.1		19.18 (14.01)	0.052		
	C	0.1		19.21 (14.07)	0.056		
RACA3 ($\tau 1 = 1$, $\tau 2 = 2$)	A	0.1	60.13	20.62 (14.06)	0.052	0.3	0.032
	B	0.1		19.90 (14.03)	0.057		
	C	0.1		19.60 (14.06)	0.048		

(continued)

Table 2 (continued)

Method	Arm	Response rate	Sample size	# of patient assigned (SD)	Pr(selected)	Degree of imbalance	Percentage of significantly imbalanced covariate
Scenario2: p1 = 0.1, p2 = 0.2, p3 = 0.3							
ER (τ1 = τ2 = 0)	A	0.1	68.48	19.89 (13.33)	0.056	0.32	0.154
	B	0.2		25.44 (12.69)	0.538		
	C	0.3		23.16 (12.10)	0.881		
RA (τ1 = 1, τ2 = 0)	A	0.1	63.71	17.16 (14.03)	0.056	0.32	0.153
	B	0.2		23.33 (13.93)	0.439		
	C	0.3		23.22 (11.97)	0.818		
CA (τ1 = 0, τ2 = 1)	A	0.1	70.51	20.28 (13.33)	0.057	0.25	0.018
	B	0.2		26.41 (12.13)	0.521		
	C	0.3		23.81 (11.93)	0.913		
RACA1	A	0.1	65.97	17.98 (14.12)	0.044	0.3	0.071
	B	0.2		24.56 (13.30)	0.511		
	C	0.3		23.43 (12.02)	0.855		
RACA2 (τ1 = τ2 = 1)	A	0.1	67.67	17.94 (14.11)	0.048	0.27	0.041
	B	0.2		24.92 (13.24)	0.525		
	C	0.3		24.81 (11.66)	0.878		
RACA3 (τ1 = 1, τ2 = 2)	A	0.1	68.49	18.89 (14.13)	0.052	0.26	0.038
	B	0.2		25.10 (13.03)	0.533		
	C	0.3		24.50 (11.78)	0.886		

without loss of statistical power. And, both designs assigned fewer patients to inferior treatment at the same level. Furthermore, the proposed RACA design is more flexible than Yuan's method. In clinical practice, with the proposed RACA design, the user can fully skew the patient allocation probability between "pure" response-adaptive and "pure" covariate-adaptive based on the trial prospective.

4 Discussion

We have developed a Bayesian RACA randomization design for multiple-arm clinical trials. We first use a prognostic score method (Yuan et al. 2011) to measure the covariate imbalance among treatment arms, then next patients' allocation probability is based on the posterior probability that assigning this patient to which treatment that minimize the covariate imbalance. We then incorporated this CA design into a RA design. The resulting design combines the advantages of CA and RA randomizations. It allocates more patients to efficacious arms, while also balancing the covariates across the treatment arms during the randomization process, as demonstrated in the simulation studies. Unlike a standard RA randomization design, our proposed design can control the covariate imbalance between the treatment arms. Consequently, the new design can help balance patient characteristics between different treatment arms, and thereby control the inflated type I error rates that occur in RA.

References

Pocock, S. J., & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 103–115.

Wei, L. J. (1978). An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association*, *73*(363), 559–563.

Atkinson, A. C. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*, *69*(1), 61–67.

Signorini, D. F., Leung, O., Simes, R. J., Beller, E., Gebski, V. J., & Callaghan, T. (1993). Dynamic balanced randomization for clinical trials. *Statistics in medicine*, *12*(24), 2343–2350.

Heritier, S., Gebski, V., & Pillai, A. (2005). Dynamic balancing randomization in controlled clinical trials. *Statistics in medicine*, *24*(24), 3729–3741.

Scott, N. W., McPherson, G. C., Ramsay, C. R., & Campbell, M. K. (2002). The method of minimization for allocation to clinical trials: a review. *Controlled clinical trials*, *23*(6), 662–674.

McEntegart, D. J. (2003). The pursuit of balance using stratified and dynamic randomization techniques: an overview. *Drug Information Journal*, *37*(3), 293–308.

Ning, J., & Huang, X. (2010). Response-adaptive randomization for clinical trials with adjustment for covariate imbalance. *Statistics in medicine*, *29*(17), 1761–1768.

Yuan, Y., Huang, X., & Liu, S. (2011). A Bayesian response-adaptive covariate-balanced randomization design with application to a leukemia clinical trial. *Statistics in medicine*, *30*(11), 1218–1229.

Lin, J., Lin, L., & Sankoh, S. (2016). A general overview of adaptive randomization design for clinical trials. *Journal of Biometrics & Biostatistics, 7*, 1–6.

Part III
Dose Ranging Studies in Clinical Trials

Sample Size Allocation in a Dose-Ranging Trial Combined with PoC

Qiqi Deng and Naitee Ting

Abstract In recent years, pharmaceutical industry has experienced many challenges in discovering and developing new drugs, including long clinical development timelines with significant investment risks. In response, many sponsors are working to speed up the clinical development process. One strategy is to combine the Proof of Concept (PoC) and the dose-ranging clinical studies into a single trial at the early Phase II development. One important question in designing this trial is how to calculate the sample size for such a study. In most of the early Phase II development programs, the budget concerns and ethical concerns may limit the total sample size for the trial. This manuscript discusses various ways of allocating the sample size to each treatment group, under a given total sample size; as well as the performance of different contrast test for PoC.

Keywords Proof of concept • Dose ranging • Trend test • Gatekeeping • Emax

1 Background

In drug development process, a candidate compound needs to go through a stringent series of testing for toxicity, pharmacokinetics, efficacy, and safety before it can be released to the market for patient use. The process involves different phases of nonclinical and clinical trials on animals, healthy volunteers and the patient population with the target disease. In most cases, the so-called phase II features a learning period when the patient population is first exposed to the test therapy for evidence of clinical benefit and risk; it is also a period to explore and recommend the commercial doses to be tested in large late phase (phase III) confirmatory clinical trials. Studies that focus on the former are commonly referred to as Proof-of-Concept (PoC) studies, while those addressing the latter are called dose-ranging studies. PoC is usually faster and cheaper as it only requires a well-tolerated dose

Q. Deng • N. Ting (✉)
Biostatistics and Data Sciences, Boehringer-Ingelheim Pharmaceuticals, Inc., 900 Ridgebury Road, P.O. Box 368, Ridgefield, Connecticut 06877-0368, USA
e-mail: naitee.ting@boehringer-ingelheim.com

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_7

of test therapy plus a placebo control group. A dose-ranging study, however, needs multiple doses of the test therapy to characterize the dose–response relationship. Therefore, a classical phase II development program usually consists of small scale PoC trials followed by moderately sized dose-ranging studies.

Sometimes it is desirable to combine the PoC and the dose-ranging studies into one single clinical trial. The advantage of such a design is to first make a "Go/No Go" decision based on PoC. If the decision is "Go", then the same study would sequentially provide dose-ranging information to help design the next study. When well designed and analyzed, the combined study saves development time by providing a range of efficacious doses going forward. The disadvantage is more investment before the concept is proven. In case the test drug does not work, this translates into larger sunk costs.

In the design of a first dose ranging clinical trial, the major challenge is that there are too many unknowns. In fact, there is very little information regarding product activities at various doses. In order to design this study, some assumptions are necessary. In this manuscript, two fundamental assumptions are needed –

1. The MTD obtained from previous studies is correct; and
2. The underlying efficacy dose–response relationship is monotonic, at least between the range of placebo and MTD.

From practical experiences, any additional assumptions about the shape of the underlying dose–response relationship could be potentially misleading (even though these two simple assumptions were not met in some practical situations). The fact is that without any one of the two assumptions, such a design is not possible. Furthermore, any additional assumption could potentially lead to very expensive failures.

In practice, for the first dose ranging study design, it is more important to cover a wide dose range, than simply adding more doses to a narrow range of doses. Dose range for a given study is defined as the ratio of the highest dose to the lowest dose. For example, a clinical trial with placebo, 20 mg, 40 mg and 80 mg of test doses, the dose range is 4 (80 divided by 20). Another trial with placebo, 0.1 mg, 1 mg, and 10 mg, the dose range is 100. Basically speaking, a trial with 4–5 test doses, plus a placebo control will deliver a good understanding of where the test medication is most active, if the dose range is wide enough, and the dose spacing is reasonable. Hamlett et al (2002) proposed to use a binary dose spacing (BDS) for dose allocation. Over the years, BDS has been successfully applied in many dose ranging designs [e.g., Ting et al. (2015); Wang and Ting (2012)].

The binary dose spacing (BDS) dose allocation is an intuitive, model independent proposal of selecting doses in designing dose-ranging clinical trials. In order to determine the doses using BDS approach, it is assumed that the maximum tolerable dose (MTD) is T, thus the design space is [0,T]. Without loss of generality, T can be taken as 1. It is assumed that the dose response relationship is monotonic. It is further assumed that the number of dose groups is known. Given this setting, doses are chosen from the interval [0, 1]. Suppose we want a design with three treatment groups, including the placebo, a low dose and a high dose. The placebo dose is taken to be zero. The challenge then is to select a low dose and a high dose, keeping in mind that we may not want a high dose too close to 1 (the assumed MTD). An intuitive approach might be to split this interval into half, giving the two intervals [0, 1/2] and (1/2, 1] and select a dose in each of these intervals. A natural choice is to select the midpoint of each interval (split each interval into half), giving the test doses, $x_1 = 1/2^2$ and $x_2 = 3/2^2$. Note that if we choose doses in the upper end (greater than 1/2 and approaching 1) then these doses tend to be toxic and is generally not of primary interest in the dose-selection process. On the other hand, activities of the lower doses are very important information for drug development. Hence the basic idea for BDS is to search for lower end of the dose range.

In another case, suppose we want to consider a design with four treatment groups including placebo, low, medium and high doses. In order to avoid selecting too high a dose which can be toxic, we may want to keep $3/2^2$ as the high dose, that is, leave the interval (1/2, 1] unchanged. Since we want to use of some low doses to help identify the MinED, we can divide the lower interval [0, 1/2] into half, giving the new intervals $[0, 1/2^2]$ and $(1/2^2, 1/2]$. We then select the midpoints of these two intervals respectively—split these two intervals into halves—giving the three test doses $x_1 = 1/2^3$, $x_2 = 3/2^3$, and $x_3 = 3/2^2$. We continue in this fashion by splitting the lower interval into half and taking the midpoint of each interval, until all the doses are allocated.

The basic concept of BDS is similar to the application of log scale dose assignment. The main idea is to define a wide dose range, and to allow very low doses can be studied in an early Phase II design. Hence in this manuscript, when three test doses are used, these doses are selected as 30 mg, 10 mg, and 3 mg. When four doses are used, they are 30 mg, 10 mg, 3 mg, and 1 mg.

2 Number of Doses and Control Groups

As discussed in the previous section, a wide dose range is critical in the phase II dose-ranging studies. Once MTD is estimated from earlier studies and a wide dose range is considered, the natural questions to think about is how many doses should be included in the study, whether a placebo control group is sufficient or an active control group is also needed.

A single test dose plus a placebo arm design may be able to demonstrate the proof-of-concept, but it cannot adequately characterize the dose–response relationship. Any attempt to interpolate a dose–response relationship between the placebo and the test dose would require very strong assumptions, which often times are not realistic. Therefore, a dose-ranging study typically needs several test dose levels. Generally speaking, it would be desirable to have more than two test doses plus one placebo arm in exploring the dose–response relationship and estimating the difference of the test doses verses placebo. One very commonly used design is a four group study with placebo, low dose, medium dose and high dose groups of the drug candidate under development. Some authors suggest that more doses could help (Krams et al. 2003)—that is, in the first dose-ranging study, adding more doses to the study design. However, the number of doses to be used in a dose-ranging trial is usually limited by practical and logistic considerations, e.g., available formulations of test doses, dosing frequencies, convenience for outpatients to use on their own or blinding complexity, etc. In some situations, an active control is also employed in a dose-ranging trial. Given these many treatment groups already included in a study, it may not be practical to add very many test doses into the same study.

In an early Phase II dose-ranging trial design, typically the total sample size is limited depending on budget, ethical considerations, availability of patients and difficulty in recruitment. Under this circumstance, more treatment groups in a design imply fewer patients per treatment group. A smaller sample size of each treatment group provides a less precise estimate of treatment effect. Hence blindly adding dose groups to a study design does not necessarily make the design to deliver more informative results. A well-thought-through design strategy is needed before the number of dose groups can be determined.

In our opinion, for the first dose ranging study design, it is more important to cover a wide dose range, than simply adding more doses to a narrow range of doses. This is both practical, and scientifically sound. In practice, a trial with 4–5 test doses, plus a placebo control will deliver a good understanding of where the test medication is most active, if the dose range is wide enough, and the dose spacing is reasonable. Some simulation studies (Yuan and Ting 2014) suggest that among the number of treatment arms (4, 5, 6, or 7, placebo arm included), it looks like three test doses (the 4-arm design) could be insufficient at some situations. The performance increases when more than three doses are studied. On the other hand, six test doses (the 7-arm design) may not necessarily deliver better results than the four test doses (the 5-arm design) or five test doses (the 6-arm design) comparisons. When the total sample size is fixed, the 7-arm design offers a smaller sample size per group, and the precision would be sacrificed. Hence from a practical point of view, a dose ranging design including 4–5 test doses (in addition to placebo) which cover a wide dose range may be very useful in designing the first dose ranging clinical trial.

3 Was the Concept Proven?

As discussed earlier, in certain cases, it is desirable to combine the PoC and the dose-ranging studies into one single clinical trial. It is important to realize that the nature of PoC is a confirmatory practice, and the corresponding statistical procedure is hypothesis testing. On the other hand, the nature of dose ranging is an exploratory practice—in fact, the project team is using this study to learn about efficacy and safety of each dose. A "dose–response relationship" can also be estimated. Meanwhile, other characteristics of a dose–response curve can also be estimated—such as minimum or maximum effective doses, safety profiles, or other parameters of interest. The key point is that information regarding to dose ranging is obtained as a learning process, various properties are observed, but they are not necessarily confirmatory features.

Therefore, the primary hypothesis should be tested for PoC in order to help with a “Go/No Go” decision. All confirmatory considerations such as alpha, power, and sample size should be focused on this primary hypothesis test. Although in some situations there could be some interests to confirm a particular dose that is significantly different from placebo, or to establish a secondary endpoint, these objectives should not interfere with the primary goal of making a “Go/No Go” decision. On this basis, all of the alpha proposed for this combined PoC and dose-ranging design should be allocated to the PoC hypothesis test. Even if there is a need to spend some alpha for secondary objectives, the study design should allow those alpha only be allocated after concept is proven. Following this thinking process, the ideal statistical hypothesis for such a study design should be a single degree of freedom test, and that the entire experimentwise alpha should be devoted to this PoC hypothesis.

Examples of combined PoC and dose-ranging designs can be found in Ting et al. (2015) and Wang and Ting (2012). In these articles, the proposed PoC can be achieved using a trend test. For example, in a four-group design with placebo, low dose, medium dose and high dose, the PoC can be only based on high dose vs placebo (let μ_L be the mean response of low dose, μ_M be the mean response for the medium dose, and μ_H be the mean response of high dose). Because a four group (three doses plus placebo) is a popular design, a set of contrasts based on four groups is described below:

$$H_0 : \mu_H = \mu_P \quad \text{vs} \quad H_1 : \mu_H > \mu_P \tag{A}$$

Or the contrast can be a trend test which is based on all doses (Wang and Ting 2012; Ting et al. 2015), assuming monotonic dose–response relationship. Table 1 lists the coefficients for trend tests under a variety number of treatment groups. The contrast for a four group design is –

$$H_0 : -3\mu_P - \mu_L + \mu_M + 3\mu_H = 0 \quad \text{vs} \quad H_1 : -3\mu_P - \mu_L + \mu_M + 3\mu_H > 0 \tag{B}$$

Or assuming only the high dose is effective –

$$H_0 : -\mu_P - \mu_L - \mu_M + 3\mu_H = 0 \quad \text{vs} \quad H_1 : -\mu_P - \mu_L - \mu_M + 3\mu_H > 0 \tag{C}$$

Table 1 Coefficients to be used in contrast for the trend test

Number of doses plus placebo	Coefficients						
	Placebo	Lowest dose	Doses increase from left to right				Highest dose
Two doses	−1	0					1
Three doses	−3	−1	1				3
Four doses	−2	−1	0	1			2
Five doses	−5	−3	−1	1	3		5
Six doses	−3	−2	−1	0	1	2	3

Or assuming high dose and median dose are equally effective, while low dose is like placebo –

$$\mathrm{H_0}: -\mu_\mathrm{P} - \mu_\mathrm{L} + \mu_\mathrm{M} + \mu_\mathrm{H} = 0 \quad \text{vs} \quad \mathrm{H_1}: -\mu_\mathrm{P} – \mu_\mathrm{L} + \mu_\mathrm{M} + \mu_\mathrm{H} > 0 \qquad \text{(D)}$$

Other possible PoC contrast may also include (assuming all doses are equally effective)

$$\mathrm{H_0}: -3\mu_\mathrm{P} + \mu_\mathrm{L} + \mu_\mathrm{M} + \mu_\mathrm{H} = 0 \quad \text{vs} \quad \mathrm{H_1}: -3\mu_\mathrm{P} + \mu_\mathrm{L} + \mu_\mathrm{M} + \mu_\mathrm{H} > 0 \quad \text{(E)}$$

As can be found from the above examples, if a contrast (that includes more than two treatment groups) is used for the PoC hypothesis, then there could be many possible ways of writing such a contrast.

4 Comparison of Power

In this section, we compare the power of five tests discussed above in Sect. 3. Chang and Chow (2006) indicated that for a dose ranging study with k arms, if we assume μ_i is the population mean for group i. The proof-of-concept can be tested using the following contrast test:

$$H_0 : L(\mu) = \sum_{i=0}^{k} c_i\mu_i = 0 \quad H_a : L(\mu) = \sum_{i=0}^{k} c_i\mu_i = \varepsilon$$

where $\sum_{i=0}^{k} c_i = 0.$

And power of the test is

$$1 - \beta = \Phi\left(\frac{\in}{\sigma}\sqrt{\frac{n}{\sum_{i=0}^{k} c_i^2/f_i}}\right)$$

Where Φ is CDF of a standard normal distribution, σ is the population standard deviation, n is the total sample size of the study and f_i is the sample size fraction for the ith group. For example, for a study with $n = 60$ and $f = (1/3, 1/6, 1/6, 1/3)$, we will allocation 20 subjects on the first and the fourth group and 10 subjects on the second and the third group respectively

Given each approach of assessing PoC described in Sect. 3, a variety of sample size-allocation can be considered in a four treatment group design (three test doses and placebo), e.g. 2:1:1:2 allocation. In a five treatment group design (four test doses and placebo), an allocation of 3:2:2:2:3, or other proposals can also be candidates for sample sizes. In other words, to allow the placebo and the high dose with more patients, and fewer patients assigned to doses in between.

Under fixed total sample size, with these five PoC contrasts, equal or unequal sample size allocation, power of PoC test is used as the metric for comparison. Because power comparisons can be evaluated analytically, there is no need to perform simulations in these comparisons.

4.1 Four-Arm Study

In this section, we compare power for a study with a total sample size of 60 patients allocated to four groups: High (30 mg), Median (10 mg), Low dose (3 mg) of test drug and placebo. These comparisons are performed under five scenarios with different dose response relationships as illustrated in Table 2 and Fig. 1.

Figure 2 shows the power of PoC versus total sample size, assuming the five different dose response shapes given in Table 2 and Fig. 1. In each plot, A,B,C,D,E represent the five different test as shown in Table 2. The solid line is equal allocation of 1:1:1:1, and the dashed line is unequal allocation of 2:1:1:2. The three points represent total sample size of 48, 60 and 72. The plot is generated using R program. The power for total sample size of 60 is also shown as in Table 3.

Table 2 Mean response for the four arms assuming a common standard deviation of 1

No.	Shape	μ_0 (placebo)	μ_1 (low)	μ_2 (median)	μ_3 (high)
1	Linear	0.15	0.24	0.45	1.05
2	Step	0.15	0.6	0.6	1.05
3	Quadratic	0.15	0.6	1.05	0.9
4	Convex	0.15	0.15	0.15	0.9
5	Concave	0.15	0.9	0.9	0.9

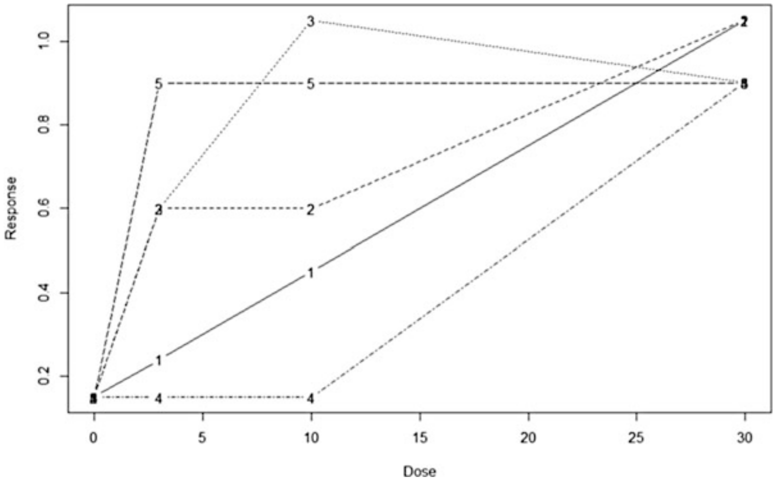


Fig. 1 Shapes of dose–response relationship evaluated under a four-arm design

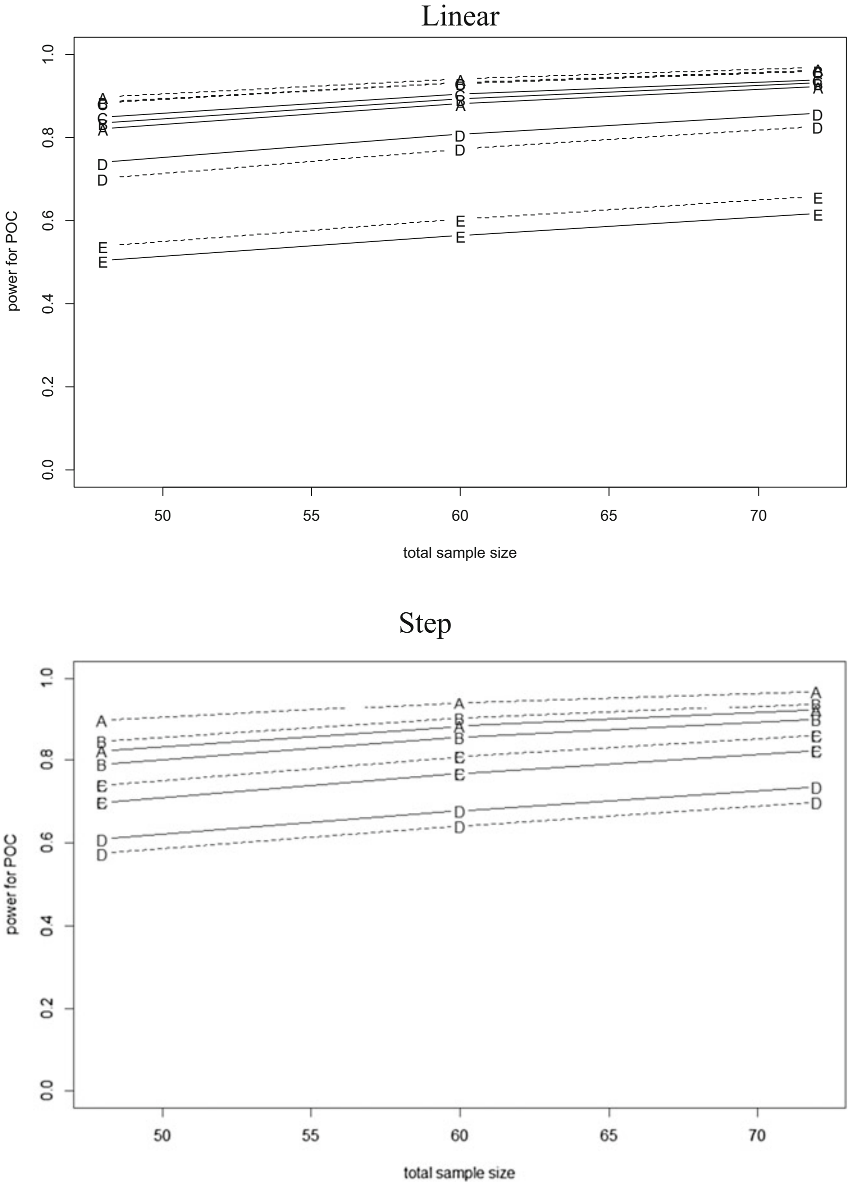
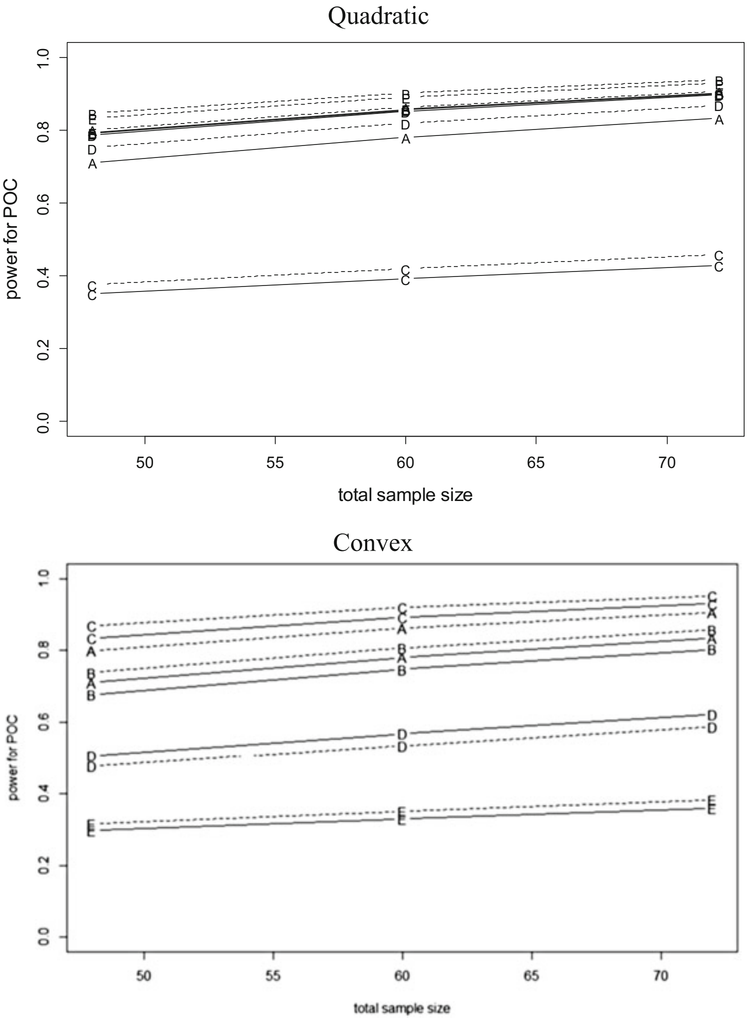


Fig. 2 Power of PoC versus total sample size, assuming different dose response shape: from *top* to *bottom* are Linear, step, quadratic, convex and concave. The *solid line* is equal allocation of 1:1:1:1, and the *dashed line* is unequal allocation of 2:1:1:2


Quadratic

Convex

Fig. 2 (continued)

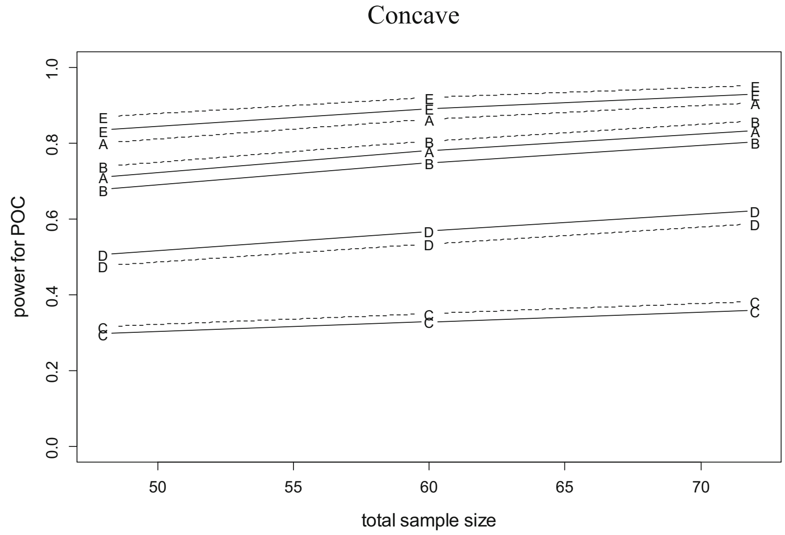


Fig. 2 (continued)

Table 3 Power for a trial with 60 patients in total, one-sided alpha = 0.1

	Method	Linear	Step	Quadratic	Convex	Concave
1:1:1:1	A: High vs PBO (−1,0,0,1)	0.88	0.88	0.78	0.78	0.78
	B: Trend Test (−3, −1, 1, 3)	0.89	0.85	0.85	0.75	0.75
	C: High vs Median/Low/PBO (−1,−1,−1,3)	0.90	0.77	0.39	0.89	0.33
	D: High/Median vs Low/PBO (−1,−1,1,1)	0.81	0.68	0.85	0.57	0.57
	E: High/Median/Low vs PBO (−3,1,1,1)	0.56	0.77	0.86	0.33	0.89
2:1:1:2	A: High vs PBO (−1,0,0,1)	0.94	0.94	0.86	0.86	0.86
	B: Trend Test (−3, −1, 1, 3)	0.93	0.90	0.90	0.81	0.81
	C: High vs Median/Low/PBO (−1,−1,−1,3)	0.93	0.81	0.42	0.92	0.35
	D: High/Median vs Low/PBO (−1,−1,1,1)	0.77	0.64	0.82	0.53	0.53
	E: High/Median/Low vs PBO (−3,1,1,1)	0.60	0.81	0.89	0.35	0.92

The methods with 2:1:1:2 (20: 10: 10: 20) allocation in general has a better performance (except for B where it is slightly worse but comparable) than 1:1:1:1 (15:15:15:15) allocation. This is not surprising given the fact that more subjects are allocated on the higher dose and the placebo where the treatment effect can be most easily differentiated.

The most powerful yet robust method is the trend test and the traditional PoC test which only use information from highest dose and placebo. Both contrasts provided more than 80 % of power consistently across different dose response shape with 2:1:1:2 allocation, and more than 75 % of power with 1:1:1:1 allocation. Although the traditional PoC test did not utilize all data in the study, the power is comparable to trend test in most of situations.

While it is not intuitive at the first glance, using average of high dose and median against the average of low dose and placebo should be avoided. It only provides around 50 % of power under a commonly seen concave response curve, and it is almost uniformly less powerful than the trend test or the traditional PoC test (except the quadratic shape under equal allocation where it is close to trend test). The loss of power is substantial under many situations.

Using high dose against average of the other three groups should be avoided by all means, since the power will be below 40 % under concave curve. Using the average of three doses from test drug against placebo could be an option, if the umbrella shape of curve is a real possibility, for example in the development of certain anti-psychotic agents. However, due to the significant loss of power to be around 30–35 % in case of a convex shape of dose–response curve, it should only be used if you have a strong confidence that the possibility of convex shape can be excluded, which we do not believe it happens very often in the first dose ranging study.

4.2 Five-Arm Study

With the learning from four arm study, we restrict our comparison for five-arm study on the three relatively robust tests given in Sect. 4.1: trend test, traditional PoC, and average the two highest vs the two lowest. The performance of other options is similar as in four-arm study. In this section, we compare power for a study with a total sample size of 60 patients again, allocated to five groups: 1, 3, 10, 30 mg dose of test drug and placebo. This comparison will be done under five scenarios with different dose response Wang and Ting (2012) used in their paper. When the total sample size is fixed, under equal allocation, the power will decrease when more arms are added into study. As a result, we increased the treatment effect in order to bring power up to the reasonable range, as illustrated in Table 4 and Fig. 3.

Table 4 Mean response for the four arms assuming a common standard deviation of 1

No.	Shape	μ_0	μ_1	μ_2	μ_3	μ_4
1	Linear	0	0.03	0.1	0.3	0.9
2	Sharp slope	0	0.01	0.12	0.6	0.65
3	Convex	0	0	0.01	0.03	0.6
4	Concave 1	0	0.12	0.29	0.57	0.8
5	Concave 2	0	0.25	0.5	0.8	0.95

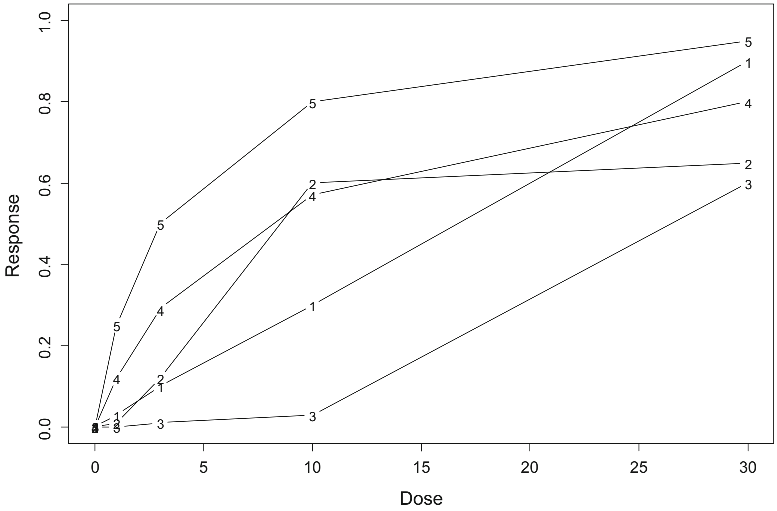


Fig. 3 Shapes of dose–response relationship evaluated under s five-arm design

Table 5 Dose–response relationships to be studied under a five-arm design

	Method	Linear	Shape slope	Convex	Concave 1	Concave 2
1:1:1:1:1	A: 30 mg vs PBO (−1,0,0,0,1)	0.82	0.62	0.57	0.75	0.85
	B: Trend Test (−2, −1, 0, 1, 2)	0.84	0.78	0.53	0.83	0.92
	D: 30/10 mg vs 1 mg/PBO (−1,−1,0, 1,1)	0.77	0.81	0.42	0.81	0.91
1.5:1:1::1:1.5	A: 30 mg vs PBO (−1,0,0,0,1)	0.88	0.69	0.64	0.82	0.91
	B: Trend Test (−2, −1, 0, 1, 2)	0.87	0.82	0.56	0.87	0.94
	D: 30/10 mg vs 1 mg/PBO (−1,−1,0, 1,1)	0.77	0.81	0.42	0.81	0.91

Again, the methods with 1.5:1:1:1:1.5 (15:10:10:10:15) allocation in general has a better performance than 1:1:1:1:1 (12:12:12:12:12) allocation (Table 5). Within more doses tested in the study, trend test starts to show better performance in terms of power than simply use high dose vs Placebo, and becomes the most powerful yet robust test. Performance of traditional PoC test is less robust in this case as in comparison with the four-arm case in previous section. All method had significant loss of power under convex curve, when the impact to the average test D is most severe.

5 Discussion

Typically a first Phase II clinical trial is designed with the primary objective of proof of concept. Such a study will help the project team in making the Go/No Go decision. Recently many sponsors tend to incorporate the PoC study with a secondary objective of exploring the dose-range. In such a design, the common assumptions include 1. MTD is correctly identified from Phase I and II. the underlying dose–response relationship is non-decreasing. Under this set up, the PoC includes two treatment groups—placebo and a high dose close to MTD, and the combined study will add a few doses between these two anchors.

In this manuscript we discussed sample size allocations in designing a combined dose-ranging and PoC clinical trial. Four treatment group (placebo plus three test doses) and five treatment group (placebo plus four test doses) designs are used in this manuscript for demonstration. In the three dose setting, the doses are 30 mg, 10 mg, and 3 mg. In the four dose setting, a 1 mg treatment group is added. Study power can be assessed analytically and hence no simulation is performed. Five sets of contrasts are considered, and a variety of underlying dose–response shapes are studied. The results indicate that the two group PoC contrasts and the trend tests provide better performances.

Based on power analyses, findings indicate that the traditional PoC comparison can still be reasonably powerful—either the design is only a two-group PoC, or a multiple group design with dose-ranging exploration. In addition, allocating more patients on the two ends increases the power when total sample size is fixed. Therefore, when the resource is limited, an imbalanced sample size allocation can be considered when designing a combined PoC and dose-ranging clinical trial. Either the two group PoC test or the trend test can be used as the primary contract in helping with a Go/No Go decision. For unbalanced designs in a four group study, for example, the ratio of 2:1:1:2 can be used. Of course other unbalance ratio with more patients allocated to the placebo control and to the highest dose, while fewer patients allocated to intermediate doses can also be considered. In case there are specific considerations, readers are encouraged to perform calculation in assessing the performance of these unbalance designs.

Although this method still leads to bigger total sample size comparing to the two-group PoC trials, the increase in sample size for the middle doses is not proportional to the increase in the number of arms, due to the lower allocation ratio, as well as the partially used information in trend test. These additional samples allocated to the middle doses have discounted contributions for PoC and Go/No Go decision. However, they can be very useful in characterizing the dose response relationship, and provide important guidance for subsequent dose finding trials. Under circumstances where clear dose response is shown, the combined study alone may justify the doses chosen to move into phase III. That potential benefit often outweighs the risk of larger upfront investment. In cases where historical information is available (for example, the response for the control arm exist in historical trials), and is deemed appropriate to be extrapolated to current study, Go/No Go decision based on Bayesian framework can be useful. Bayesian approach may further reduce the actual sample size for PoC by having virtual patients in control arm through prior. Nevertheless, to obtain information about the dose response relationship, allocating some patients to middle doses may still be helpful.

In design and analysis of dose-ranging clinical trials, there are typically hypothesis testing approaches and estimation approaches. Under the hypothesis testing frame work, alpha control is very critical, and multiple comparison adjustment would be applied. On the other hand, estimations are usually achieved by using dose–response models. In the model based approaches, often alpha control comes from testing certain parameter(s) in the given model. In general, the advantages of using dose–response models including –

1. Minimum effective dose and other target doses can be estimated;
2. Confidence intervals on these doses can be constructed;
3. All doses within the observed dose range can be studied; and
4. Avoids multiple comparison adjustment.

However, modeling approach requires additional assumptions, and some models needs more dose groups in the design. Sample size calculation could be more complicated. This manuscript takes the hypothesis testing approach and hence other than monotonicity, no additional model assumption is considered here.

References

Chang M, Chow SC. (2006) Power and sample size for dose–response studies, *Dose Finding in Drug Development,* Springer, New York, pp 220–241

Hamlett, A., N. Ting, C. Hanumara, J.S. Finman, Dose Spacing in Early Dose Response Clinical Trial Designs. Drug Information Journal (2002) Vol. 36, Issue 4, pp 855–864

Ting, N., G. Yuan, F. Beckers, Classic dose response study. Clinical trials using SAS: Classical, Adaptive, and Bayesian Methods. SAS Institute, to appear in 2015

Wang, X., N. Ting. A Proof-of-concept Clinical Trial Design Combined with Dose-Ranging Exploration, Biopharmaceutical Statistics, (2012) wileyonlinelibrary.com DOI: 10.1002/pst.1525

Krams M, KR Lees, W Hacke, AP Grieve, JM Orgogozo, GA Ford, and for the ASTIN Study Investigators. (2003). ASTIN: An adaptive dose–response study of UK-279,276 in acute ischemic stroke. Stroke. 34, 2543–254.

Yuan, G., and Ting, N., Clinical Trial Biostatistics and Biopharmaceutical Applications (2014), pp 55–74, Chapman and Hall/CRC

Personalized Effective Dose Selection in Dose Ranging Studies

Xiwen Ma, Wei Zheng, and Yuefeng Lu

Abstract We consider the problem of predicting the personalized minimum effective dose and estimating the dose-dependent optimal subgroups in dose-ranging studies. Our research is motivated by a real randomized, double-blind, placebo-controlled phase II dose-ranging study with genetic markers. One goal of the analysis is to identify subgroups with enhanced benefit/risk profiles with approriate doses and inform the study design of future phase III trials. To the best of our knowledge, this problem has not been systematically studied before. We proposed a novel framework to nonparametrically model the dose-dependent biomarker-outcome relationship and to estimate the personalized effective dose and dose-dependent optimal subgroups. Our proposed method will be useful for identifying the respondent subgroups and their accompanying doses for the future study design. We illustrate the proposed method with simulation studies. Our method compares favorably to two ad-hoc approaches.

Keywords Dose-ranging study • Personalized medicine • Individualized treatment selection • Personalized dose selection • Dose-dependent subgroup

1 Introduction

For pharmaceutical interventions, it's well known that the strategy of "one-size fits all" is hardly applicable to most common diseases. Spear, Heath-Chiozz and Huff reported that the percentage of patients for whom drugs are ineffective ranges from 38 to 75 % for several major diseases, due to the heterogeneity of patient population, complex underlying pathophysiology, and inadequate or inappropriate dosing regimens among other factors (Spear et al. 2001). With recent advances in biological science and enhanced understanding of diseases at the level of molecular biology and pathophysiology, opportunities are created to fulfill these unmet needs, often by leveraging on genetic, genomic and imaging biomarkers. The science of

X. Ma (✉) • W. Zheng • Y. Lu
Biostatistics and Programming, Sanofi, 1 The Mountain Rd., Framingham, MA 01701, USA
e-mail: xiwen.ma@sanofi.com

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_8

personalized medicine involves developing and validating evidence-based treatment algorithms to match a right patient with the right treatment, at the right dose and at the right time.

Our work is motivated by a real randomized, double-blind, placebo-controlled phase II dose-ranging study of a novel treatment for asthma patients. Biomarkers for individual patients, including genetic markers and patient baseline demographics, were collected. One study goal is to explore the relationship between the dose, biomarker and outcome and consequently identify the right patient population as well as appropriate doses for future study designs. Compared to the two-arm design, a dose-ranging study provides some unique advantages for the purpose of personalized medicine. First, dosing is an import dimension for treatment decision. For example for patients treated with the anticoagulant Warfarin, careful dose titration needs to be conducted to maintain the therapeutic target. Conversely inadequate or inappropriate dosing is thought as a major factor leading to sub-optimal clinical outcomes (Spear et al. 2001). It's plausible to postulate that the relationship between biomarkers and outcomes is dose-dependent, and therefore the optimal subgroups for the treatment are different for different doses. Secondly, for predictive biomarkers previously identified in preclinical or phase I studies, it's preferable to further demonstrate them in a dose-response fashion. By imposing the requirement of dose-response in biomarkers, we will likely reduce the chance of having spurious findings. Lastly, from the perspective of drug development, it is desirable to develop tailoring strategy before phase III confirmatory trials are carried out. Usually it's in phase II trials when for the first time the efficacy is evaluated in patients as a primary goal, thus methods for identifying target patients under such settings would be valuable.

To the best of our knowledge, the problem of personalized dose selection in dose-ranging studies has not been considered before. A number of statistical approaches have recently been proposed for personalized treatment selection and subgroup identification, all under the setting of a randomized clinical study comparing a new treatment with the standard of care (Dusseldorp and Van Mechelen 2013; Foster et al. 2011; Lipkovich et al. 2011; Loh et al. 2015; Zhao et al. 2013; Cai et al. 2011; Claggett et al. 2015; Matsouaka et al. 2014; Kang et al. 2014; Huang and Fong 2014; Zhao et al. 2012, 2015). For dose-ranging studies, some ad-hoc approaches might be considered. One approach is to only use the highest dose, under the assumption that stronger therapeutic effects are manifested at higher doses. This however only utilizes partial data and likely suffers from issues of lack of power. Another approach is to group all doses, reducing the problem to a two-arm study. A drawback of this approach is that the between-dose difference is inappropriately modeled as part of the inter-subject variability. For both approaches, important dose information is not fully utilized, which can result in incomplete or wrong conclusions, since again the optimal subgroups are likely dose-dependent.

In this paper we present a novel framework for personalized dose selection for individual patients in dose-ranging studies. The dose-biomarker-outcome relationship is estimated with nonparametric methods, with the constraint of the outcome being monotonic in dose for given biomarkers. Bootstrapping is applied to obtain confidence intervals or confidence regions for the estimates.

The paper is organized as follows. In Sect. 2, we introduce the model and present the method for estimating the dose-dependent biomarker-outcome relationship as well as estimations of the personalized effective dose and optimal subgroups. We illustrate our approach with simulation studies in Sect. 3. We conclude with a discussion in Sect. 4.

2 Methods

2.1 Context, Notation and Model

In a phase II dose-ranging study, patients are randomized into T dosing groups with doses from a dose space $\mathcal{A} = \{d_1, d_2, \ldots, d_T\}$, where $d_1 < d_2 < \cdots < d_T$. Note that $d_1 = 0$ denotes the placebo group or $d_1 > 0$ denotes the minimum dose group when there is no placebo group such as in some oncology studies. To concentrate on the main ideas, we assume $d_1 = 0$ throughout the paper but the ideas can be generalized to the case when $d_1 > 0$ (also see discussion in the last section). Let Y denote the clinical outcome and $\mathbf{x} = (x_1, x_2, \ldots, x_p)' \in \mathcal{X}$ denote biomarkers. Without loss of generality, we assume that a larger value of Y indicates a more favorable clinical outcome. A commonly adopted assumption in dose-ranging studies is that the mean clinical outcome in the overall population is monotonic in dose. It's reasonable to expect the monotonic assumption still holds for given biomarkers. Similar assumptions have been made in dose-response microarray experiments (Lin et al. 2007).

Assumption 1. *For any biomarker* $\mathbf{x}$, *one of the following conditions holds:*

$$1.\ E(Y|\mathbf{x}, d_1) \leq E(Y|\mathbf{x}, d_2) \leq \cdots \leq E(Y|\mathbf{x}, d_T),$$

$$2.\ E(Y|\mathbf{x}, d_1) \geq E(Y|\mathbf{x}, d_2) \geq \cdots \geq E(Y|\mathbf{x}, d_T).$$

When there is a strong belief that the treatment increases the efficacy, we may only impose the monotonically increasing condition. For generality, we consider both monotonically increasing and decreasing conditions in this paper.

A personalized dose can be viewed as a decision rule from the baseline biomarker space $\mathcal{X}$ to the dose space $\mathcal{A}$:

$$D : \mathcal{X} \to \mathcal{A}$$

For a given efficacy margin $\delta > 0$ and biomarkers $\mathbf{x}$, the personalized effective dose (PED) is the minimum dose at which the expected clinical outcome exceeds that of the placebo by an efficacy margin δ. More precisely, given a margin $\delta > 0$, the PED is

$$D^*(\mathbf{x}, \delta) = \min\{d_i \in \mathcal{A} : E(Y|\mathbf{x}, d_i) - E(Y|\mathbf{x}, d_1) \geq \delta\}.$$

Note that PED is only well defined under the monotonic assumption. When no dose meets the above requirement, PED is defined as $+\infty$ for increasing dose-response or $-\infty$ for decreasing dose-response for convenience.

Given an efficacy margin $\delta > 0$, the Dose-Dependent Optimal Subgroup (DDO-Subgroup) for dose d_i is defined as

$$\mathcal{S}(d_i, \delta) = \{\mathbf{x} \in \mathcal{X} : E(Y|\mathbf{x}, d_i) - E(Y|\mathbf{x}, d_1) \geq \delta\}.$$

which is the biomarker subspace on which the efficacy margin is at least δ at dose d_i. Under the monotonicity Assumption 1, it's easy to validate that the optimal subgroup for a lower dose must be a subset of the subgroup for a higher dose. In the next subsection, we propose an algorithm combining nonparametric regression and isotonic adjustment to estimate the dose-dependent biomarker-outcome relationship.

2.2 Estimation: The INIA Method

In this paper, we consider the case of the outcome y being either continuous or binary. To simplify the notation, we denote the mean response function as

$$g(\mathbf{x}, d) = E(Y|\mathbf{x}, d).$$

For continuous outcomes, we have

$$Y = g(\mathbf{x}, d) + e, \text{where } e \sim \mathrm{N}(0, \sigma^2), \tag{1}$$

and for binary outcomes, we have

$$\mathrm{P}(Y = 1|\mathbf{x}, d) = g(\mathbf{x}, d). \tag{2}$$

We assume that n_i patients are treated at the dose d_i, and denote the total sample size as n. Let $\mathbf{x}_{ij} = (x_{ij}^1, x_{ij}^2, \ldots, x_{ij}^p)$ denote the vector of biomarkers for the jth patient at the dose d_i and y_{ij} denote the patient's clinical outcome, $i = 1, \ldots, T$, $j = 1, \ldots, n_i$.

We let $g_i(\mathbf{x}) = E(Y|\mathbf{x}, d_i), i = 1, \ldots, T$, the mean response function for each dose group. We propose an approach of Iterative dose-dependent Nonparametric regression with Isotonic Adjustment (INIA) to estimate the dose-dependent mean response functions with monotonic constraints. The algorithm iteratively performs nonparametric regression independently at each dose followed with isotonic regression to ensure the validity of the monotonic constraints. The details of the algorithm are as follows:

1. Initial fitting: Fit the regression function $\hat{g}_i(\mathbf{x})$ to the data at dose d_i with a nonparametric regression method such as smoothing splines (Wang 2011; Wahba and Craven 1979), gradient boosting (Breiman 1996), or random forest (Breiman 2001), etc.
2. Isotonic adjustment: For each observation $\mathbf{x}_{ij}$, obtain its predicted outcomes at all doses $\hat{g}_k(\mathbf{x}_{ij})$ from step 1, $k = 1, \ldots, T$. The isotonic adjustment is then applied to $\hat{g}_k(\mathbf{x}_{ij})$'s at $\mathbf{x}_{ij}$:

   a. We perform isotonic regression to the T predicted values $\hat{g}_k(\mathbf{x}_{ij}), k = 1, \ldots, T$ assuming either increasing or decreasing dose-response by optimizing the following

   $$\begin{array}{l}\min_{a_1,\ldots,a_T} \sum_{k=1}^{T} \left[a_k - \hat{g}_k(\mathbf{x}_{ij})\right]^2, \\ s.t.\ \ a_1 \le a_2 \le \cdots \le a_T\end{array} \tag{3}$$

   or

   $$\begin{array}{l}\min_{a_1,\ldots,a_T} \sum_{k=1}^{T} \left[a_k - \hat{g}_k(\mathbf{x}_{ij})\right]^2. \\ s.t.\ \ a_1 \ge a_2 \ge \cdots \ge a_T\end{array} \tag{4}$$

   The Pool-Adjacent-Violators algorithm is used to obtain the solutions (de Leeuw et al. 2009).

   b. We compare the residual sum of squares from the two models (3) and (4) and choose the model with smaller errors. Denote the solution from the chosen model as $\hat{a}_k, k = 1, \ldots, T$. The predicted value at $\mathbf{x}_{ij}$ and dose d_k is then:

   $$\hat{y}_{ij}^{(k)} = \hat{a}_k. \tag{5}$$

   Thus we obtain the augmented data $\{\hat{y}_{ij}^{(1)}, \ldots, \hat{y}_{ij}^{(T)}\}$ for all doses at $\mathbf{x}_{ij}$. Note at every dose the size of augmented data is nT.
3. Refitting the augmented data: Update the estimated mean response function $\hat{g}_k(\mathbf{x})$ by fitting the augmented data $\{(\mathbf{x}_{ij}, \hat{y}_{ij}^{(k)}), i = 1, \ldots, T, j = 1, \ldots, n_i\}$ obtained in step 2. For binary outcomes, $\hat{y}_{ij}^{(k)}$'s are the predicted probabilities and our refitting procedure is similar to the quasi maximum likelihood estimate for fractional response data in Papke and Wooldridge (1996).
4. Final model: iterate between step 2 and 3 until it converges.

The plug-in estimate for the PED and the DDO-subgroup is then

$$\hat{D}^*(x, \delta) = \min\{d_i : \hat{g}_i(\mathbf{x}) - \hat{g}_1(\mathbf{x}) \ge \delta\}.$$

and

$$\hat{\mathcal{S}}(d_i, \delta) = \{\mathbf{x} : \hat{g}_i(\mathbf{x}) - \hat{g}_1(\mathbf{x}) \ge \delta\}.$$

respectively. Confidence intervals for the mean response functions and confidence regions for DDO-subgroups are constructed using bootstrapping.

3 Numerical Examples

3.1 Examples and Evaluation Criteria

We first illustrate the proposed method with examples of a single or two biomarkers. We then present a multi-marker example with simulated SNPs data. In each example, there is a placebo group, a low, median and high dose group. We compare our method to two ad-hoc approaches: Only using the highest dose (High-Only), and grouping all doses (Group-All). The estimates of PED and DDO-subgroup are evaluated with the following criteria: misclassification rate (MR), sensitivity (SEN), specificity (SPE), positive predictive value (PPV) and negative predictive value (NPV).

3.2 Single Marker Examples

The first example is simulated from the mean response function:

$$g(x,d) = \begin{cases} 0.1, & \text{if } x \le \gamma(d) \\ \sqrt{x-\gamma(d)} + 0.1, & \text{if } \gamma(d) < x \end{cases}$$

where x is uniformly distributed over $[0, 1]$ and $\gamma(d) = 0.75, 0.5$ and 0.25 for the low, median and high dose group respectively. The sample size is 100 for each group. A thousand simulations are performed for both continuous and binary outcomes. In each simulation, the efficacy margin δ is uniformly drawn from the interval $[0.1, 0.4]$.

Continuous outcomes are generated from (1) with $\sigma = 0.1$. We use the smoothing spline (Wang 2011) to estimate the mean response functions in the proposed INIA algorithm. Figure 1 shows a typical example by the proposed method including the initial fit, final estimates of mean response functions, their 95 % confidence intervals, and estimated PEDs and DDO-subgroups. Table 1 summarizes comparisons of estimated PEDs and DDO-subgroups. For PED, the “Only-High” approach has inferior specificity, PPV and MR compared to the proposed method, because the “Only-High” approach only uses the high dose and thus tends to overestimate the minimum efficacious dose for those patients whose true PED is low or median dose. Since for the “Group-All” approach dosing information is lost and all dose groups are mixed, the statistics in Table 1 are meant to be interpreted as if the assumed single treatment dose is low, median and high respectively while in fact there are three doses in the study. Compared to the proposed method, the “Group-All” approach has inferior specificity, PPV and MR for the low dose group because some patients whose PED is in fact median or high are thought to achieve the efficacy margin at the mistakenly assumed low dose; it also has inferior sensitivity,

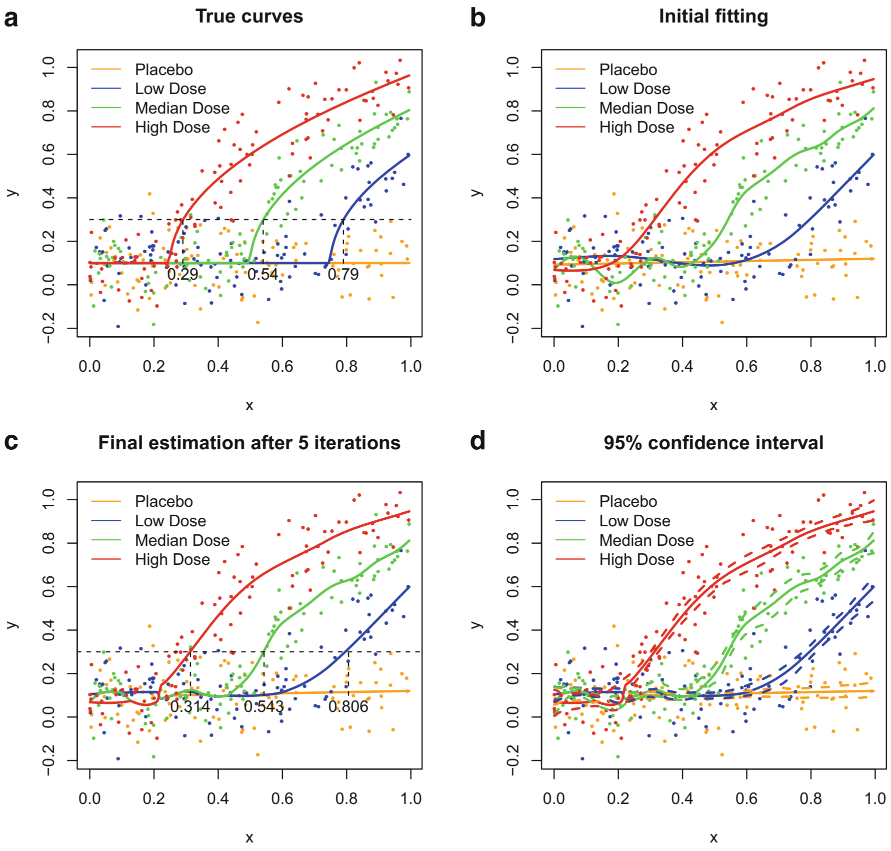


Fig. 1 An example with single-marker simulation with continuous outcomes. (**a**) true mean response functions and observations (**b**) initial fit (**c**) final fit after 5 iterations where the horizontal dash line indicates the efficacy margin and the vertical dash lines indicate the cut-off points for DDO-subgroups (**d**) 95 % confidence intervals by bootstrapping

NPV and MR for the high dose group because it is less sensitive in identifying those patients whose outcome is only slightly higher than the efficacy margin at the high dose likely due to the inflated between-subject errors by mixing all doses. Overall the proposed method has superior performance by utilizing all the data and modeling the dose-response.

Binary outcomes are generated from (2). We apply the gradient boosting (Friedman 2000) to fit the mean response functions in the INIA algorithm. A typical example is shown in Fig. 2. The summary statistics for comparing PED and DDO-subgroup estimates are presented in Table 2. Similar observations can be made as those for continuous outcomes.

Table 1 Comparison of the PED and DDO-subgroup estimates for the single marker example with continuous outcomes

		Only-High			Group-All			Our method		
	Stat	Low	Median	High	Low	Median	High	Low	Median	High
PED	SEN	0	0	0.95				0.91	0.94	0.93
	SPE	1	1	0.42				0.99	0.97	0.98
	PPV	0	0	0.36				0.97	0.93	0.94
	NPV	0.82	0.75	0.97				0.98	0.98	0.98
	MR	0.18	0.25	0.45				0.02	0.04	0.03
Subgroup	SEN			0.98	1	0.95	0.66	0.91	0.98	0.99
	SPE			0.99	0.66	0.92	1	0.99	0.98	0.97
	PPV			0.99	0.4	0.93	1	0.97	0.98	0.99
	NPV			0.97	1	0.97	0.6	0.98	0.99	0.98
	MR			0.02	0.27	0.059	0.23	0.02	0.02	0.02

Note that PED can't be estimated by the "Group-All" method and DDO-subgroup can't be estimated by the "Only-High" method

3.3 A Two-Marker Example

In this example, we simulate data from the following mean response function:

$$g(x,d) = \begin{cases} 0, & \text{if } \sqrt{x_1} + \sqrt{x_2} \le \gamma(d) \\ \log(\sqrt{x_1} + \sqrt{x_2} - \gamma(d) + 1), & \text{if } \sqrt{x_1} + \sqrt{x_2} > \gamma(d) \end{cases}$$

where x_1 and x_2 are independently drawn from Uniform[0, 1] and the cutoff point $\gamma(d) = 2, 1.25, 1, 0.75$ is for the placebo, low, median and high dose respectively. The sample size is 100 for each group. A thousand simulations are performed for both continuous and binary outcomes. In each simulation, the efficacy margin δ is uniformly drawn from the interval $[0.1, 0.4]$.

Continuous outcomes are generated from (1) with $\sigma = 0.06$. We apply nonparametric kernel regression (Hayfield and Racine 2008) to fit the mean response functions in the INIA algorithm. Figure 3 demonstrates our proposed method. Figure 4 plots the contours for DDO-subgroups for an example, where the blue, green and red solid line defines the boundary of the true DDO-subgroup for the low, median and high dose respectively with the efficacy margin $\delta = 0.3$, and the dotted lines are their estimated counterparts. The estimates of PED and DDO-subgroups are summarized in Table 3. We have similar observations as those for the single-marker examples.

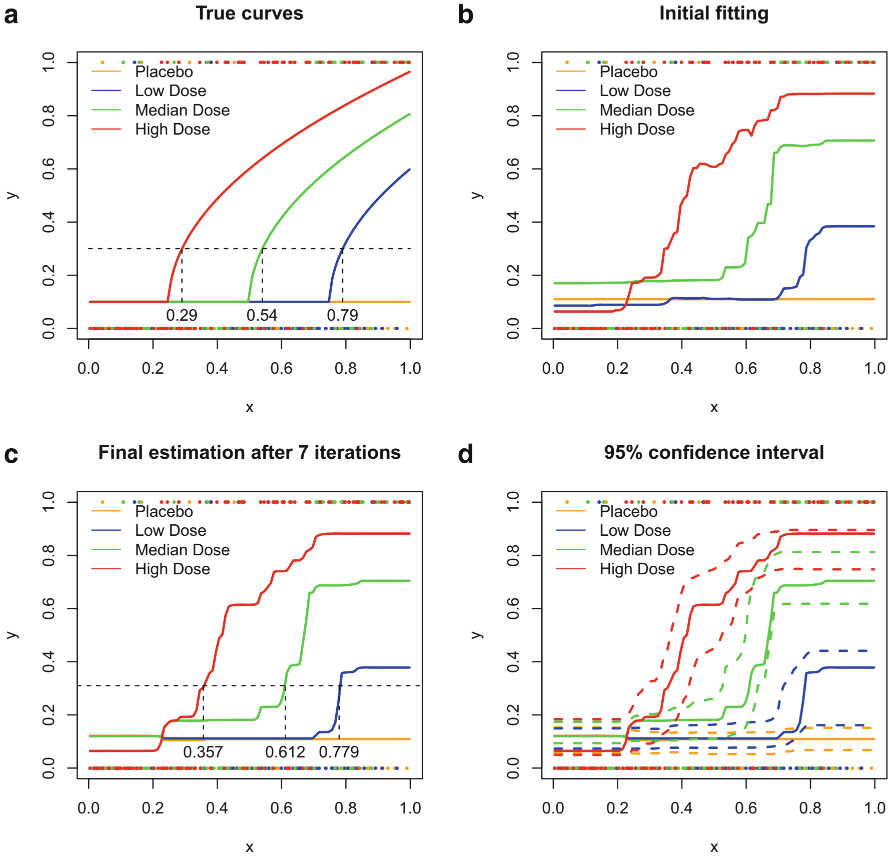


Fig. 2 An example with single-marker simulation with binary outcome. (**a**) true mean response functions and observations (**b**) initial fit (**c**) final fit after 5 iterations where the *horizontal dash line* indicates the efficacy margin and the *vertical dash lines* indicate the cut-off points for DDO-subgroups (**d**) 95 % confidence intervals by bootstrapping

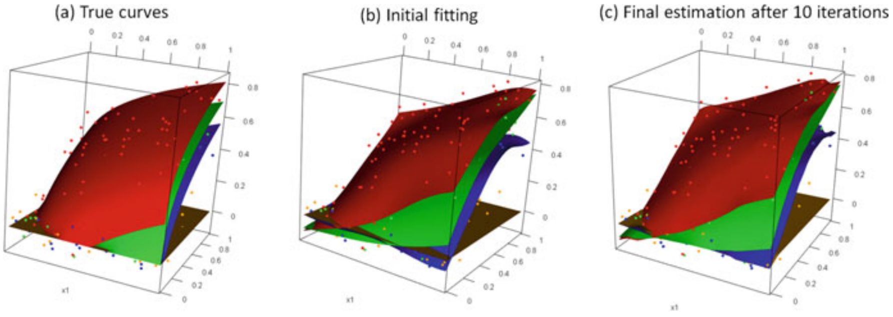
3.4 A Multi-Marker Example

We simulate a dose-ranging study with genotyping markers to emulate the real Phase-II trial aforementioned in the Introduction. The continuous outcome is simulated to represent the outcome of the lung function test, and the binary outcome is simulated to represent the event of exacerbation. A hundred single nucleotide polymorphisms (SNPs) are simulated under the Hardy-Weinberg equilibrium with minor allele frequency ranging between 0.01 and 0.5. The first 10 SNPs are prognostic markers independent of the treatment and the next 10 SNPs are dose-dependent predictive markers. Specifically data are generated from the following mean response function:

Table 2 Comparison of the PED and DDO-subgroup estimates for the single marker example with binary outcomes

		Only-High			Group-All			Our method		
	Stat	Low	Median	High	Low	Median	High	Low	Median	High
PED	SEN	0	0	0.82				0.92	0.71	0.78
	SPE	1	1	0.40				0.96	0.95	0.92
	PPV	0	0	0.31				0.86	0.88	0.79
	NPV	0.82	0.75	0.89				0.98	0.91	0.93
	MR	0.18	0.25	0.49				0.05	0.11	0.12
Subgroup	SEN			0.93	1	0.91	0.63	0.92	0.90	0.95
	SPE			0.94	0.68	0.92	0.99	0.95	0.96	0.93
	PPV			0.98	0.43	0.93	1	0.86	0.96	0.97
	NPV			0.89	1	0.95	0.59	0.98	0.94	0.92
	MR			0.07	0.26	0.077	0.25	0.05	0.06	0.06

Note that PED can't be estimated by the "Group-All" method and DDO-subgroup can't be estimated by the "Only-High" method


(a) True curves

(b) Initial fitting

(c) Final estimation after 10 iterations

Fig. 3 An example with two-marker simulation with continuous outcomes. *Brown*=placebo, *blue*=low dose, *green*=median dose and *red*=high dose. (**a**) true response functions and observations (**b**) initial fitting (**c**) Final fit after 10 iterations

$$g(x, d) = \frac{1}{c}\left(\sum_{i=1}^{10} x_i + d\sum_{i=11}^{20} x_i\right). \tag{6}$$

where $x_i = 0, 1$ or 2 is the number of minor alleles for the ith SNP and d is the dose. We use d =0, 5, 10 and 20, for the placebo, low, median and high dose respectively, with the scaling parameter $c = 1$ for the continuous outcome and $c = 150$ for the binary outcome. The continuous outcome is generated from (1) with $\sigma = 2$ and the binary outcome is generated from (2). In each simulation, we generate a training dataset and a test dataset, both having 400 samples (100 samples for each group). For each simulation, the efficacy margin δ is drawn uniformly from the interval $[1, 50]$ for the continuous outcome, and from the interval $[0.05, 0.5]$ for the binary outcome. We apply our method to the training data and evaluate the

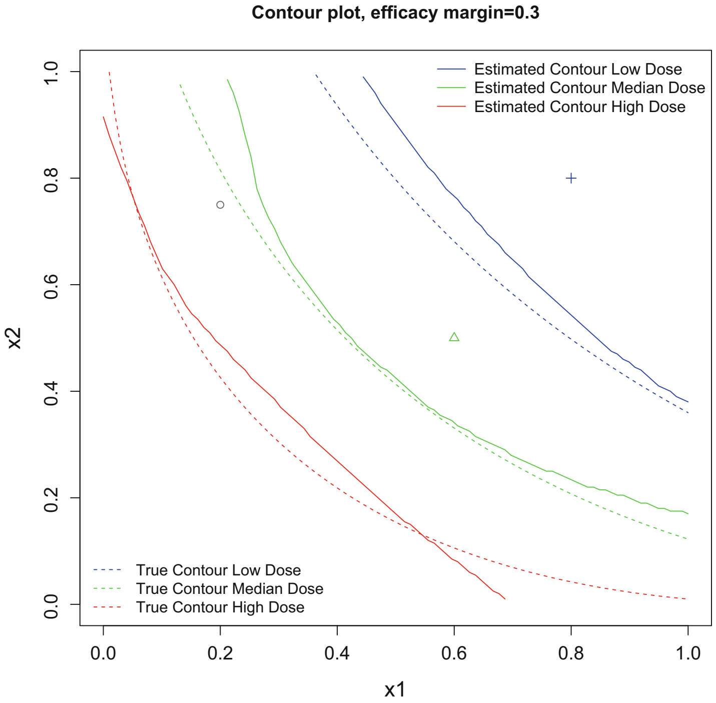


Fig. 4 An example with two-marker simulation with continuous outcomes for estimating DDO-subgroups. The *blue, green* and *red solid line* defines the boundary of the true DDO-subgroup for the low, median and high dose respectively with the efficacy margin $\delta = 0.3$; the *dotted lines* are their estimated counterparts

performance on the test data using 1000 simulations. Gradient boosting is used to fit the mean response functions.

Table 4 provides the summary statistics for the estimates of PED and DDO-subgroup for continuous outcomes, and Table 5 provides those for binary outcomes. We have similar observations as in the previous examples. The proposed method compares favorably to the two ad-hoc approaches.

4 Discussion

In this paper, we present a novel framework to model the dose-response biomarker-outcome relationship in dose-ranging studies and apply it to estimate the personalized effective dose and dose-dependent optimal subgroups. To the best of our knowledge, this problem has not been systematically studied before. Our proposed method can be useful for identifying respondent subgroups and their accompanying

Table 3 Comparison of the PED and DDO-subgroup estimates for the two-marker example with continuous outcomes

		Only-High			Group-All			Our method		
	Stat	Low	Median	High	Low	Median	High	Low	Median	High
PED	SEN	0	0	0.90				0.90	0.87	0.86
	SPE	1	1	0.27				0.99	0.96	0.94
	PPV	0	0	0.27				0.98	0.88	0.82
	NPV	0.7	0.74	0.91				0.97	0.95	0.96
	MR	0.3	0.26	0.59				0.03	0.07	0.08
Subgroup	SEN			0.97	1	0.94	0.69	0.90	0.94	0.97
	SPE			0.93	0.63	0.94	1	0.99	0.97	0.91
	PPV			0.98	0.53	0.97	1	0.98	0.98	0.98
	NPV			0.91	1	0.94	0.48	0.97	0.94	0.91
	MR			0.04	0.25	0.049	0.24	0.03	0.04	0.04

Note that PED can't be estimated by the "Group-All" method and DDO-subgroup can't be estimated by the "Only-High" method

Table 4 Comparison of the PED and DDO-subgroup estimates for the multi-marker example with continuous outcomes

		Only-High			Group-All			Our method		
	Stat	Low	Median	High	Low	Median	High	Low	Median	High
PED	SEN	0	0	0.79				0.70	0.47	0.81
	SPE	0.95	1	0.03				0.59	0.65	0.88
	PPV	0	0	0.13				0.54	0.42	0.33
	NPV	0.52	0.65	0.45				0.57	0.77	0.90
	MR	0.48	0.35	0.86				0.13	0.27	0.16
Subgroup	SEN			0.99	0.82	0.96	0.93	0.89	0.98	0.99
	SPE			0.27	0.096	0.19	0.31	0.59	0.49	0.35
	PPV			0.98	0.49	0.58	0.78	0.64	0.89	0.88
	NPV			0.39	0.53	0.40	0.16	0.57	0.48	0.39
	MR			0.03	0.44	0.13	0.077	0.13	0.09	0.03

Note that PED can't be estimated by the "Group-All" method and DDO-subgroup can't be estimated by the "Only-High" method

doses for future study designs. We demonstrate our method through simulation studies and show our method compares favorably to two ad-hoc approaches.

In this paper the efficacy margin is defined in the form of differences in mean response functions, e.g. risk difference for binary outcomes. The main ideas can be readily extended for other definitions of the efficacy margin, such as odds ratio. Also we have assumed throughout the paper there is a placebo group in the study. For some oncology studies, there is no placebo group for ethical reasons. For those cases, we can use the lowest dose in lieu of placebo in the proposed method, and define the efficacy margin using an absolute value instead of the change from

Table 5 Comparison of the PED and DDO-subgroup estimates for the multi-marker example with binary outcomes

		Only-High			Group-All			Our Method		
	Stat	Low	Median	High	Low	Median	High	Low	Median	High
PED	SEN	0	0	0.00				0.26	0.21	0.18
	SPE	0.56	1	0.05				0.42	0.62	0.67
	PPV	0	0	0.00				0.35	0.20	0.12
	NPV	0.0042	1	0.25				0.49	0.80	0.80
	MR	0.95	0	0.95				0.65	0.38	0.33
Subgroup	SEN			0.95	0.43	0.58	0.82	0.46	0.64	0.97
	SPE			0.11	0.18	0.14	0.12	0.42	0.31	0.13
	PPV			0.99	0.65	0.75	0.76	0.75	0.72	0.99
	NPV			0.09	0.004	0.003	0.003	0.12	0.11	0.09
	MR			0.04	0.28	0.30	0.36	0.20	0.21	0.03

Note that PED can't be estimated by the "Group-All" method and DDO-subgroup can't be estimated by the "Only-High" method

placebo. Note when there is no placebo group, generally we can't estimate the prognostic effects and may not be able to estimate some predictive effects.

We want to discuss a limitation of our approach. The phase II dose-ranging studies in some therapeutic areas can have a small sample size. For small sample-size studies, the performance of the proposed method can quickly deteriorate with the increasing number of noisy biomarkers. Filtering, pre-screening or unsupervised learning are generally recommended to reduce the number of biomarkers before applying the proposed method. Impact of sample size and biomarker design on the proposed method will be analyzed in the future research.

We haven't considered the biomarker selection or investigated the theoretical properties of the proposed method. Both will be interesting topics for future research. Another interesting topic would be approximating DDO-subgroups with more regular boundaries such as cubes defined by trees or affine hyperplanes defined by linear combination of biomarkers.

References

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman. Bias, variance, and arcing classifiers. Technical report, 1996.

Tianxi Cai, Lu Tian, Peggy H. Wong, and L. J. Wei. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282, 2011.

Brian Claggett, Lu Tian, Davide Castagno, and L. J. Wei. Treatment selections using riskbenefit profiles based on data from comparative randomized clinical trials with multiple endpoints. *Biostatistics*, 16(1):60–72, 2015.

Jan de Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009.

Elise Dusseldorp and Iven Van Mechelen. Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. *Statistics in Medicine*, 2013.

Fared C. Foster, Jeremy M.G. Taylor, and Stephen J. Ruberg. Sugroups identification from randomized clinical trial data. *Statistics in Medicine*, 2011.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

Tristen Hayfield and Jeffrey S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.

Ying Huang and Youyi Fong. Identifying optimal biomarker combinations for treatment selection via a robust kernel method. *Biometrics*, 70(4):891–901, 2014.

Chaeryon Kang, Holly Janes, and Ying Huang. Combining biomarkers to optimize patient treatment recommendations. *Biometrics*, 70(3):695–707, 2014.

Dan Lin, Ziv Shkedy, Dani Yekutieli, Tomasz Buryzykowski, Hinrich W.H. Gohlmann, An De Bondt, Tim Perera, Tamara Geerts, and Luc Bijnens. Testing for trends in dose-response microarray experiments: A comparison of several testing procedures, multiplicity and resampling-based inference. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

Ilya Lipkovich, Alex Dmitrienko, Jonathan Denne, and Gregory Enas. Subgroup identification based on differential effect seawrch - a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 2011.

W.-Y. Loh, X. He, and M. Man. A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, 2015.

Roland A. Matsouaka, Junlong Li, and Tianxi Cai. Evaluating marker-guided treatment selection strategies. *Biometrics*, 70:489–499, 2014.

Leslie E. Papke and Jeffrey M. Wooldridge. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11:619–632, 1996.

Biran B Spear, Margo Heath-Chiozzi, and Jeffrey Huff. Clinical application of pharmacogenetics. *Trends in Molecular Medicine*, 7:201–204, 2001.

Grace Wahba and P. Craven. Smoothing noisy data with spline functions. *Numerische Mathematik*, (31):337–403, 1979.

Yuedong Wang. *Smoothing splines: methods and applications*. Monographs on Statistics and Applied Probability 121. CRC Press, Boca Raton, 2011.

Lihui Zhao, Lu Tian, Tianxi Cai, Brian Claggett, and L. J. Wei. Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association*, 108(502):527–539, 2013.

Y. Q. Zhao, D. Zeng, E. B. Laber, R. Song, M. Yuan, and M. R. Kosorok. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102:151–168, 2015.

Y. Q. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107: 1106–1118, 2012.

Part IV
Innovative Clinical Trial Designs and Analysis

Evaluation of Consistency Requirements in Multi-Regional Clinical Trials with Different Endpoints

Zhaoyang Teng, Jianchang Lin, and Bin Zhang

Abstract In recent years, there is an increasing trend to conduct multi-regional clinical trials (MRCT) for drug development in Pharmaceuticals industry. A carefully designed MRCT could be used in supporting the new drug's approval in different regions simultaneously. The primary objective of an MRCT is to investigate the drug's overall efficacy across regions while also assessing the drug's performance in some specific regions. In order to claim the study drug's efficacy and get drug approval in some specific region(s), the local regulatory authority may require the sponsors to provide evidence of consistency in the treatment effect between the overall patient population and the local region. Usually, the regional specific consistency requirement needs to be pre-specified before the study conduct and the consistency in treatment effect between the region(s) of interest and overall population will be evaluated at the final analysis. In this paper, we evaluate the consistency requirements in multi-regional clinical trials for different endpoints, i.e., continuous, binary and survival endpoints. We also compare the different consistency requirements of the same endpoint/measurement if multiple consistency requirements are enforced and our recommendations for each endpoint/measurement will be made based on the comprehensive consideration.

Keywords Multi-regional clinical trials • Consistency requirement • Assurance probability

1 Introduction

With the increasing of globalization of drug development, more and more pharmaceutical industries are conducting multi-regional clinical trials (MRCT) to support a new drug's efficacy across different populations. A carefully designed MRCT

Z. Teng (✉) • J. Lin
Takeda Pharmaceuticals, Cambridge, MA, USA
e-mail: zhaoyang.teng2@takeda.com

B. Zhang
Genocea Biosciences, Cambridge, MA, USA

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_9

could be used to support the new drug's approval in different regions/countries simultaneously. It should at least provide a pathway for a regulatory agency to ensure the drug's safety and efficacy based not only on the overall patient population but also the local region.

One of the statistical challenges in conducting MRCT is to ensure that overall efficacy can be adequately preserved in regions of interested. To evaluate the possibility of applying the overall results in a MRCT to the specific regions of interest, the Japanese Ministry of Health, Labour and Welfare (MHLW 2007) proposed two methods for determining the needed number of Japanese subjects for establishing consistency of the treatment effect between the Japanese patients and others:

- Method 1, the sample size needed for the Japanese patients in a MRCT was to satisfy

$$P\left(D_i/D > \pi_i\right) > 1 - \beta^{'},$$

where D_i and D are the estimated treatment effects from the group of Japanese patients and the entire patient population, respectively, π_i is the effect retention rate ≥ 0.5 and $\beta^{'}$ is the type II error rate ≤ 0.2.

- Method 2, the planned sample size for the Japanese group in a MRCT was to satisfy

$$P\left(D_1 > 0, D_2 > 0, \ldots, D_s > 0\right) > 1 - \beta^{'},$$

where D_i represents the observed treatment effect for region i, i = 1, ... ,s. Here the s is to denote the number of regions.

Based on the Japanese MHLW guidance, a couple of statistical methods were proposed to apply the overall results of the MRCT to the specific region. On the basis of Method 2, Kawai et al. (2008) proposed an approach to partition the total sample size to the individual regions to ensure a high probability of observing a consistent trend if the treatment effect is positive and uniform across the regions. On the basis of Method 1, Quan et al. (2010a, b) discussed the sample size requirement for normal, binary and survival endpoint. Quan et al. (2010a, b) proposed five definitions of consistency, and calculated the probability of consistency for different configurations of sample size allocations and true treatment effects in individual regions. Ko et al. (2010) proposed four criteria to determine whether the treatment is effective in a specific region given the overall result is significant at the α level. Tsong et al. (2012) then proposed an approach to control the type I error rate of a specific region adjusted by the regional sample size. In particular, what they proposed was to determine the sample size of a MRCT to accommodate the overall type I error rate as well as the regional specific type I error rate. Chen et al. (2012) proposed two conditional decision rules for regional approval, where sample size determination and the relationship between the two rules were also discussed.

Quan et al. (2012) proposed the empirical shrinkage estimation approach based on the random effect model to assess the consistency of treatment effect across regions, which presumably could help obtain better consistency compared to the fixed effect model. Tsou et al. (2012) also proposed another consistency criterion to examine whether the overall results can be applied to all participating regions, sample size requirement were also discussed.

It should be noted that most available consistency requirement for MRCT were proposed for continuous endpoint. The consistency requirement for binary and survival endpoints have not been well discussed in the literature. In this paper, we evaluated the impact on consistency requirements in MRCT for different endpoints: continuous, binary and survival. Additionally, we will compare the different consistency requirements of the same endpoint/measurement if multiple consistency requirements are available and give the general recommendations for each endpoint/measurement based on the practical consideration.

We organize this paper as follows. In Sect. 2, we present the various statistical models available for MRCT. In Sect. 3, we discuss the consistency requirement for different endpoints; the comparisons of different consistency requirement for the same endpoint/measurement are also discussed in this section. In Sect. 4, we use an example from an oncology trial to illustrate the practical application of the proposed methods. Finally, we present summary and discussion about the use of our methods in Sect. 5.

2 Fixed and Random Effect Models

Assume μ_i is the true treatment effect for regions i, σ_i^2 is the variance of treatment effect in region i who are in either test or control group. Although σ_i^2 can be different across regions, in this paper, we assume that σ_i^2 are the same for all regions, namely, $\sigma_i^2 = \sigma^2$, $i = 1, \ldots, s$. Denote D_i the estimated treatment effect for region i, N_i the number of patients/events in each arm of region i, N the total number of patients/events per arm in an MRCT, and $f_i = N_i/N$ the proportion of patients/events in regions i to the total number of patients/events. If the overall treatment effect for the entire MRCT is the weighted combination of regional treatment effect and the sample size/event proportion is considered as weight which is also commonly used, then the overall treatment effect is $D = \sum_{i=1}^{s} (f_i D_i)$, where s is the number of regions in an MRCT. Suppose the endpoint follows a normal distribution, i.e.

$$D_i \Big| \mu_i \sim N\left(\mu_i, \frac{2\sigma^2}{N_i}\right)$$

Then

$$D\Big|\mu_i,\ i = 1\ldots,s \sim N\ \left(\mu, \frac{2\sigma^2}{N}\right),\ where\ \mu = \sum_{i=1}^{s} f_i\mu_i$$

In the conventional clinical trial, the total sample size N for a trial is usually determined to detect the expected treatment difference μ at the desired one-sided significance level α and $1 - \beta$ power. Using a two sample z-test, it is clear that

$$N = 2\left[\frac{\left(z_{1-\alpha} + z_{1-\beta}\right)\sigma}{\mu}\right]^2 \tag{1}$$

Where $z_{1-\alpha}$ and $z_{1-\beta}$ are the $(1 - \alpha)$ th and $(1 - \beta)$ th percentile of the standard normal distribution, respectively.

The model described above is the fixed effect model where μ_i*'s* are fixed parameters. Besides that, random effect model proposed by Hung et al. (2010), et.al is another useful model which could be applied to MRCT.

In the random effect model, the μ_i's are no longer fixed parameters but random variables with the same distribution, e.g.

$$\mu_i \sim N\left(\delta,\ \pi^2\right)$$

and unconditionally

$$D_i \sim N\left(\delta,\ \pi^2 + \frac{2\sigma^2}{N_i}\right)$$

In this paper, we focus on the fixed effect model, and all the methodologies proposed in this paper could be extended to random effect model and other model such as discrete random effects model proposed by Lan and Pinheiro (2012).

3 Consistency Requirement of Different Endpoints

In this paper, we focus on the consistency in treatment effect between one specific region and overall results. The simultaneous consistency assessment of different endpoints for all regions will be discussed in a separate paper from authors. In practice, the consistency in treatment effect will be evaluated only when the overall efficacy result is statistically significant. Assuming that MHLW Method 1 is the consistency requirement, we introduce the definition of assurance probability which is the probability of region i satisfying the regional requirement given the overall efficacy, i.e.

$$AP_i = P_\mu\left(D_i/D > \pi_i\Big|Z > z_{1-\alpha}\right)$$

where Z is the test statistic of the overall efficacy i.e. $Z = D/std(D) = D\sqrt{N}/\sqrt{2\sigma^2}$ and $z_{1-\alpha}$ is the $(1-\alpha)$ th percentile of the standard normal distribution. In this definition, we are interested in the chance of demonstrating the consistency trend in treatment effect by assuming that a significant global result is obtained already. Additionally, the success for a region could be defined as the probability of region i satisfying the regional requirement and observing the positive overall result, i.e.

$$S_i = P_\mu \left(D_i/D > \pi_i,\ Z > z_{1-\alpha} \right)$$

Numerically, the success for a region equals to the production of assurance probability of this region and the overall power. The assurance probability highly depends on the sample size/event proportion of this region and would not increase too much with total sample size increase. However, the success of this region will be improved because of the increased overall power when the sample size proportion of this region is fixed.

3.1 Continuous Endpoint

Most of methodologies are less challenging to be implemented in continuous endpoint compared to binary and survival endpoints. Partially because of this reason most of the published consistency requirements in literature were proposed for continues endpoint. As mentioned in Chen et al. (2012), two qualitative methods were usually considered to assess the consistency in treatment effect between each region and the entire trial. The first is to evaluate if the estimated regional treatment effect preserves some proportion of overall treatment effect, i.e. $D_i > \pi_i D$, which utilizes the spirit of Method 1 in Japanese guidance. Here the value of π_i should be pre-specified. The second is to test if the treatment effect based on the samples from local region is statistically significant at level α_i, i.e. $D_i > z_{1-\alpha_i}\sqrt{2\sigma^2/N_i}$. Tsong et al. (2012) proposed to determine the regional type I error rate α_i adjusted for the regional sample size. Recently, Teng and Chang (2016) proposed a unified consistency requirement which generalizes the two above consistency requirements. This unified consistency requirement is to test whether the "true" regional treatment effect preserve some proportion of the "true" overall treatment effect at significance level α_i, we can use hypothesis test

$$H_0 : \mu_i \le \pi_i \mu \ \ versus \ \ H_a : \mu_i > \pi_i \mu$$

Denote Z_i as the test statistics of the hypothesis test above, where $Z_i = (D_i - \pi_i D) / std (D_i - \pi_i D)$, and then $Z_i > z_{1-\alpha_i}$ is the unified consistency requirement. When $\alpha_i = 0.5$, $Z_i > z_{1-\alpha_i}$ is reduced to $D_i > \pi_i D$; when $\pi_i = 0$, $Z_i > z_{1-\alpha_i}$ is reduced to $D_i > z_{1-\alpha_i}\sqrt{2\sigma^2/n_i}$. Two parameters π_i and α_i are involved in this consistency requirement, which make it more commonly feasible.

For the illustration purpose, we utilize the first type of consistency requirement for binary and survival endpoint.

3.2 Binary Endpoint

Suppose $n_i^l \sim B\left(N_i,\ p_i^l\right)$ is the number of events from the lth treatment group in region i, and $\hat{p}_i^l = \frac{n_i^l}{N_i}$ is the estimated of the event rate p_i^l, $i = 1, \cdots, s$ and $l = t, c$ which represent treatment group and control group, respectively. We assume the higher event rate is, the better the result is, e.g. response rate in oncology trials. There are three major measurements of treatment effect for binary endpoint, i.e.

Risk different: $\hat{rd}_i = \hat{p}_i^t - \hat{p}_i^c$
Relative risk: $\hat{rr}_i = \hat{p}_i^t / \hat{p}_i^c$
Odds ratio: $\hat{or}_i = \frac{\hat{p}_i^t\left(1-\hat{p}_i^c\right)}{\hat{p}_i^c\left(1-\hat{p}_i^t\right)}$

Similar to continuous endpoint, the overall treatment effect is estimated from the weighted average of regional treatment, i.e. $\hat{b} = \sum_{i=1}^{s}\left(f_i \hat{b}_i\right)$, where $\hat{b} = \hat{rd}, \hat{rr},\ \hat{or}$ and $\hat{b}_i = \hat{rd}_i, \hat{rr}_i, \hat{or}_i$.

3.2.1 Risk Difference

Suppose the regional risk difference $\hat{rd}_i = \hat{p}_i^t - \hat{p}_i^c$ follows a normal distribution asymptotically,

$$\hat{rd}_i \sim N\left(p_i^t - p_i^c, \frac{p_i^t\left(1-p_i^t\right) + p_i^c\left(1-p_i^c\right)}{N_i}\right)$$

Then the distribution of the overall risk difference can be derived as

$$\hat{rd} \sim N\left(\sum_{i=1}^{s}\frac{N_i}{N} rd_i, \sum_{i=1}^{s}\frac{N_i\left(p_i^t\left(1-p_i^t\right) + p_i^c\left(1-p_i^c\right)\right)}{N^2}\right)$$

If the first type of regional consistency requirement is imposed for regional approval, the criterion below needs to be satisfied in order to claim the regional efficacy:

$$\hat{rd}_i > \pi_i \hat{rd}$$

Accordingly, the assurance probability can be expressed as follows:

$$AP_i = P_\mu\left(\hat{rd}_i > \pi_i \hat{rd} \middle| Z > z_{1-\alpha}\right)$$

The asymptotic joint distribution of the overall treatment effect and the regional treatment effect of region i is

$$\begin{pmatrix} \hat{rd}_i \\ \hat{rd} \end{pmatrix} \sim N\left(\begin{pmatrix} rd_i \\ rd \end{pmatrix}, \begin{pmatrix} \frac{p_i^t(1-p_i^t)+p_i^c(1-p_i^c)}{N_i} & \frac{p_i^t(1-p_i^t)+p_i^c(1-p_i^c)}{N} \\ \frac{p_i^t(1-p_i^t)+p_i^c(1-p_i^c)}{N} & \sum_{i=1}^{s} \frac{N_i(p_i^t(1-p_i^t)+p_i^c(1-p_i^c))}{N^2} \end{pmatrix}\right)$$

3.2.2 Relative Risk

Similar to risk difference, the asymptotic distributions of the overall treatment effect and the regional treatment effect of region i for log of relative risk are

$$\log(\hat{r}r_i) \sim N\left(\log\left(\frac{p_i^t}{p_i^c}\right), \frac{\left(1-p_i^t\right)/p_i^t + \left(1-p_i^c\right)/p_i^c}{N_i}\right)$$

$$\log(\hat{r}r) \sim N\left(\sum_{i=1}^{s} \frac{N_i}{N} \log(rr_i), \sum_{i=1}^{s} \frac{N_i\left(\left(1-p_i^t\right)/p_i^t + \left(1-p_i^c\right)/p_i^c\right)}{N^2}\right)$$

Therefore, the joint distribution of the overall treatment effect and the regional treatment effect of region i is asymptotically

$$\begin{pmatrix} \log(\hat{r}r_i) \\ \log(\hat{r}r) \end{pmatrix} \sim N\left(\begin{pmatrix} \log(rr_i) \\ \log(rr) \end{pmatrix}, \begin{pmatrix} \frac{(1-p_i^t)/p_i^t+(1-p_i^c)/p_i^c}{N_i} & \frac{(1-p_i^t)/p_i^t+(1-p_i^c)/p_i^c}{N} \\ \frac{(1-p_i^t)/p_i^t+(1-p_i^c)/p_i^c}{N} & \sum_{i=1}^{s} \frac{N_i((1-p_i^t)/p_i^t+(1-p_i^c)/p_i^c)}{N^2} \end{pmatrix}\right)$$

Two consistency requirements could be considered for relative risk, the first one is the risk reduction; the second one is based on the log scale of relative risk.

Risk reduction: $(\hat{r}r_i - 1) > \pi_i (\hat{r}r - 1)$
Log scale of relative risk: $\log(\hat{r}r_i) > \pi_i \log(\hat{r}r)\ \Delta \hat{r}r_i > \hat{r}r^{\pi_i}$

When comparing the two consistency requirements above, it is easy to prove that the consistency requirement of risk reduction is more stringent than log scale of relative risk with the same value of π_i between 0 and 1 when the overall result is positive, i.e. $\hat{r}r > 1$. The detailed mathematical proofs are given in the Appendix. Thus, we may prefer to use the log scale of relative risk as consistency requirement since it is more powerful to preserve the consistency of treatment between the local region and overall result.

In terms of odds ratio, the distributions of the overall treatment effect, the regional treatment effect of region i and the joint distribution of them can be derived similarly. The consistency requirement proposed for relative risk could be applied to odds ratio directly and the property of the two consistency requirement still holds.

3.3 Survival Endpoint

As mentioned in Quan et al. (2010a, b), the proportional hazards model is usually considered for survival endpoint.

$$\lambda_1(t) = \lambda_0(t)e^{\gamma}$$

where $\lambda_1(t)$ and $\lambda_0(t)$ are the hazard function for treatment and control group, respectively; e^{γ} is the hazard ratio between treatment and control group. If the power is calculated based on log-rank test, the total number of events (E) needed from the two groups combined to preserve $1-\beta$ power at two-sided α level could be calculated as follows:

$$E = \frac{4\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\gamma^2}$$

The distributions of regional treatment effect $\hat{\gamma}_i$ is asymptotically

$$\hat{\gamma}_i \sim N\left(\gamma_i, \frac{4}{E_i}\right)$$

where E_i is the total number of events in region i. Assuming that the overall treatment effect for the entire MRCT is the weighted combination of regional treatment effect and the event proportion is considered as weight for survival endpoint, the overall treatment effect is $\hat{\gamma} = \sum_{i=1}^{s} \left(\frac{E_i}{E}\hat{\gamma}_i\right)$. The distribution of the overall treatment effect is asymptotically

$$\hat{\gamma} \sim N\left(\sum_{i=1}^{s} \frac{E_i}{E}\gamma_i, \frac{4}{E}\right)$$

Therefore, the joint distribution of the overall treatment effect and the regional treatment effect of region i is asymptotically

$$\begin{pmatrix} \hat{\gamma}_i \\ \hat{\gamma} \end{pmatrix} \sim N\left(\begin{pmatrix} \gamma_i \\ \gamma \end{pmatrix}, \begin{pmatrix} \frac{4}{E_i} & \frac{4}{E} \\ \frac{4}{E} & \frac{4}{E} \end{pmatrix}\right)$$

Similar to relative risk for binary endpoint, two consistency requirements could be considered for survival endpoint, the first one is hazard reduction; the second one is based on the log scale of hazard ratio.

Hazard reduction: $\left(1 - e^{\hat{\gamma}_i}\right) > \pi_i \left(1 - e^{\hat{\gamma}}\right) \Delta \left(1 - \hat{HR}_i\right) > \pi_i \left(1 - \hat{HR}\right)$

Log scale of hazard ratio: $\hat{\gamma}_i < \pi_i \hat{\gamma} \Delta \hat{HR}_i < \hat{HR}^{\pi_i}$

When comparing the two consistency requirements above, it can be proven that the consistency requirement of log scale of hazard ratio is more stringent than hazard reduction with the same value of π_i between 0 and 1 when the overall result is positive, i.e. $\hat{HR} < 1$. The detailed mathematical proofs are given in the Appendix. Thus, different from relative risk for binary endpoint we may prefer to use the hazard reduction as consistency requirement for survival endpoint. For example, if the hazard ratio for the overall result is 0.7 and $\pi_i = 0.5$. In order to claim consistency in treatment effect between region i and overall efficacy result, the estimated hazard ratio for region i should be less than 0.837 based on log scale of hazard ratio. A relative loose threshold i.e. 0.85 needs to be met when hazard reduction is considered as the consistency requirement with the same value of $\pi_i = 0.5$.

4 Example

Assume that we conduct a randomized, double-blinded, placebo-control MRCT in patients with relapsed and/or refractory multiple myeloma (RRMM) from different regions including Japan, the European Union, the United States and China. The primary objective of this MRCT is to test whether the additional of the new drug to the standard of care is better than the standard of care alone. The primary endpoint is progression free survival (PFS), the key secondary endpoint is overall survival (OS), other secondary endpoint includes overall response rate (ORR), and the very good response rate (VGPR) and complete response rate (CR), etc. PFS is generally acceptable as the primary endpoint for drug approval and registration. However, OS benefit is also very important if it can be demonstrated in the same trial even though it is difficult to demonstrate the OS benefit due to crossover issues in practice. This study is planned to power both PFS and OS. In order to obtain 80 % power to detect the hazard ratio of 0.77 for OS, 700 patients will be enrolled to obtain 486 OS events. In order to control the family-wise type I error rate, sequential testing procedure will be implemented for PFS and OS, which means OS will be tested only when we pass the test of PFS. PFS and OS follow their own alpha spending functions with O'Brien-Fleming boundaries. Three interim analyses and one final analysis are planned. The first interim analysis is for both of PFS and OS. The second interim analysis is the final analysis for PFS and the second interim analysis for OS. The third and the final analysis are for OS only.

Only 40 Japanese patients were planned for this MRCT based on the regulatory interactions. In terms of consistency requirements for regional approval, we only need to show the positive trend for Japan population due to the small number of patients enrolled from this region, i.e. $\pi_i = 0$ after negotiation with regulatory agency. First, we would like to evaluate the probability of demonstrating this specific consistency requirement in the primary endpoint (PFS) and key secondary endpoint (OS). We assume the hazard ratios of PFS and OS are the same for Japan patients and overall patients, i.e. $HR_{PFS} = 0.73, HR_{OS} = 0.77$. The alpha spending structure

Table 1 The alpha spending structure, number of events from Japan and global study and the results (assurance probability (AP), overall power and regional success of Japan) at each interim analysis (IA) for PFS and OS

		α	E_J	E_O	AP_J (%)	$Power_O$ (%)	S_J (%)
PFS	First IA	0.0163	10	262	74.1	55.8	41.3
	Second IA	0.0451	18	365	77	84.2	64.8
OS	First IA	0.0001	5	154	78.3	1.17	0.9
	Second IA	0.0018	10	222	78.3	12	9.4
	Third IA	0.0112	16	322	77.1	42.4	32.7
	Final	0.0462	26	486	77.3	81.2	62.8

and the projected number of events from Japanese patients and overall population at each interim are summarized in the first part of Table 1; the result including assurance probability; overall power and regional success of Japan at each interim are summarized in the second part of this table.

As demonstrated in Table 1, the assurance probability does not increase too much, but the regional success for Japan change from 41 % to 65 % as the overall power increasing for PFS. The same story to OS, there is no big change for assurance probability from first interim analysis to the final analysis, but the regional success of Japan changes from 0.9 % to 62.8 % as the overall power increasing. This example also verify the point we made in the previous section, the assurance probability of a region would not increase too much with total sample size/events increase, but the success of this region will be improved because of the increased overall power when the sample size/event proportion of this region is not change too much. Thus, the most efficient way to increase the assurance probability of a region is to raise the sample size/event proportion.

In most cases, it will take longer time to collect the survival data compared to binary data in the same clinical trials, especially for the effective drug. Take the MRCT we mentioned above as an example, the median time to response are 1.9 months vs. 1.1 months, whereas the median PFS are 15 months vs. 20.6 months for control and treatment group, respectively. Thus, we might already collect the adequate data of ORR in the first few months, but it will much longer time to collect the PFS data. If the ORR benefit could be translated to the PFS benefit, we may consider demonstrating the consistency in ORR to satisfy the regional requirement at the early stage of MRCT when the data of PFS is not mature. For example, we can define the consistency in ORR as the follows since we do not test the efficacy in ORR:

$$Consistency_J = P_\mu \left(\hat{rr}_J > 1,\ \hat{rr} > 1\right)$$

where $\hat{rr}_J$ and $\hat{rr}$ are the relative risks of ORR for Japanese and overall population, respectively. Even though we may not obtain a higher chance of demonstrating the consistency in treatment effect by using the endpoint with shorter time to accumulate adequate data since the assurance probability highly depends on the sample size/events proportion of this region, it will provide us more accurate and reliable result in the early stage of MRCT. For example, there are only 5 PFS events out of 40 Japanese patients at the first interim analysis which happened about 5 months after last patient in (LPI), but most of patients have reached their best response at this moment. Thus, it is a reasonable approach to demonstrate the consistency in ORR instead of PFS at the first interim analysis if this proposal is also acceptable by regulatory agency.

5 Conclusion

In general, there are two main objectives in conducting a MRCT; the first one is to evaluate the overall efficacy of the study drug in the global setting and the second is to avoid unfavorable results in some specific regions for regional approval in the case when the global results are favorable. Thus, the specific consistency requirement from the region(s) of interest may need to be satisfied in order to have the regional approval in addition to the overall efficacy. In this paper, we evaluate the consistency requirement for different endpoints. We discuss three types of consistency requirement for continuous endpoint, the first is to evaluate if the estimated regional treatment effect preserves some proportion, i.e. π_i of overall treatment effect; the second is to test if the treatment effect based on the samples from local region is statistically significant at level α_i, the third is to test whether the "true" regional treatment effect preserve some proportion, i.e. π_i of the "true" overall treatment effect at significance level $\mathrm{\alpha_i}$. Two parameters π_i and α_i are involved in the third consistency requirement making it more commonly feasible. In this paper, we focus on the first type of consistency requirement for binary and survival endpoints even thought the third type of consistency requirement is more preferable in general. However, the methodology proposed for binary and survival endpoints could be easily extended to the second and third type of consistency requirement.

One or more than one consistency requirements were proposed for the three measurements of binary endpoint and survival endpoint. We summarize the consistency requirement for different endpoint in Table 2, and also indicate which consistency requirement is preferable in general for each endpoint/measurement. One should have better chance to demonstrate the consistency in treatment effect by imposing the recommended consistency requirement when the true underlying treatment effects are in fact consistent between the local region and overall population.

Additionally, we may consider demonstrating the consistency in secondary endpoints that are correlated with our primary endpoint but take shortening time to collect adequate data if the data of the primary endpoint is not relatively mature at the early stage of MRCT due to the late enrollment of targeted regions, e.g. the first interim analysis of a MRCT with multiple interim analysis planned. However, it is always necessary to get agreement with each local regulator on which endpoint we can use to demonstrate the consistency and what the specific

Table 2 Summary of consistency requirement for different endpoint

Endpoint	Regional requirement	Preferred in general
Continues	Preserve π_i proportion of overall efficacy: $D_i > \pi_i D$	
	Regional significance at α_i: $D_i/std\,(D_i) > z_{1-\alpha_i}$	
	Unified consistency requirement: $(D_i - \pi_i D)\,/std\,(D_i - \pi_i D) > z_{1-\alpha_i}$	Yes
Binary: risk difference	Preserve π_i proportion of overall efficacy: $\hat{rd}_i > \pi_i \hat{rd}$	Yes
Binary: relative risk	Risk reduction: $(\hat{rr}_i - 1) > \pi_i\,(\hat{rr} - 1)$	
	Log scale of relative risk: $\log(\hat{rr}_i) > \pi_i \log(\hat{rr})$	Yes
Binary: odds ratio	Similar to relative risk	
Survival	Hazard reduction: $\left(1 - e^{\hat{\gamma}_i}\right) > \pi_i\left(1 - e^{\hat{\gamma}}\right)$	Yes
	Log scale of hazard ratio: $\hat{\gamma}_i < \pi_i \hat{\gamma}$	

consistency requirement they prefer. Then we allocate the total sample size to each region appropriately to guarantee certain level of regional success besides the overall success by incorporating the consistency requirement from all the regions of interest.

A.1 Appendix

A.1.1 Comparison of the Two Consistency Requirements for Relative Risk

Two consistency requirements for relative risk:

Risk reduction: $(\hat{rr}_i - 1) > \pi_i\,(\hat{rr} - 1)$
Log scale of relative risk: $\log(\hat{rr}_i) > \pi_i \log(\hat{rr})$

The proof below shows the consistency requirement of risk reduction is more stringent than log scale of relative risk with the same value of π_i when the overall result is positive.

Proof: For risk reduction $(\hat{rr}_i - 1) > \pi_i\,(\hat{rr} - 1)\,\Delta \hat{rr}_i > \pi_i\,(\hat{rr} - 1) + 1$
And for log scale of relative risk $\log(\hat{rr}_i) > \pi_i \log(\hat{rr})\,\Delta \hat{rr}_i > \hat{rr}^{\pi_i}$
Let$f(\hat{rr}) = \hat{rr}^{\pi_i} - (\pi_i\,(\hat{rr} - 1) + 1)$, then $\frac{df(\hat{rr})}{d\hat{rr}} = \pi_i\left(\hat{rr}^{\pi_i - 1} - 1\right)$

We usually further evaluate the consistency requirement given we observe a positive overall result, thus it is reasonable to assume $\hat{rr} > 1$. Since $0 \le \pi_i \le 1$, then

$$\frac{df\left(\hat{rr}\right)}{d\hat{rr}} \le 0$$

Since $f(1) = 0$, then $f\left(\hat{rr}\right) = \hat{rr}^{\pi_i} - \left(\pi_i\left(\hat{rr} - 1\right) + 1\right) \le 0$ when $\hat{rr} > 1$, Thus

$$\hat{rr}^{\pi_i} \le \left(\pi_i\left(\hat{rr} - 1\right) + 1\right)$$

In summary, we conclude the consistency requirement of risk reduction is more stringent than log scale of relative risk with the same value of π_i when the overall result is positive.

A.1.2 Comparison of the Two Consistency Requirements for Hazard Ratio

Two consistency requirements for hazard ratio:

Hazard reduction: $\left(1 - e^{\hat{\gamma}_i}\right) > \pi_i\left(1 - e^{\hat{\gamma}}\right)\Delta\left(1 - \hat{HR}_i\right) > \pi_i\left(1 - \hat{HR}\right)$

Log scale of hazard ratio: $\hat{\gamma}_i < \pi_i\hat{\gamma}\Delta\hat{HR}_i < \hat{HR}^{\pi_i}$

The proof below shows the consistency requirement of log scale of hazard ratio is more stringent than hazard reduction with the same value of π_i between 0 and 1 when the overall result is positive.

Proof: For risk reduction $\left(1 - \hat{HR}_i\right) > \pi_i\left(1 - \hat{HR}\right)\Delta\hat{HR}_i < 1 - \pi_i\left(1 - \hat{HR}\right)$

And for log scale of relative risk $\hat{HR}_i < \hat{HR}^{\pi_i}$

Follow the similar steps of proof for relative risk we can show that $1 - \pi_i\left(1 - \hat{HR}\right) \ge \hat{HR}^{\pi_i}$ with the same value of π_i between 0 and 1 when the overall result is positive, i.e. $\hat{HR} < 1$. Therefore, we conclude the log scale of hazard ratio is the more stringent consistency requirement than hazard reduction with the same value of π_i between 0 and 1 when the overall result is positive.

References

Chen, X. Y., Lu, N., Nair, R., Xu, Y. L., Kang, C. L., Huang, Q., Li, N., Chen, H. Z. (2012). Decision Rules and Associated Sample Size Planning for Regional Approval Utilizing Multiregional Clinical Trials. Journal of Biopharmaceutical Statistics 22:1001–1018.

Hung, H. M. J., Wang, S. J., O'Neill, R. T. (2010). Consideration of regional difference in design and analysis of multi-regional trials. Pharmaceutical Statistics 9:173–178.

Lan, K. K. G., Pinheiro, J. (2012). Combined estimation of treatment effects under a discrete random effects model. Statistics in Biosciences 14:235–244

Kawai, N., Stein, C., Komiyama, O., Ii, Y. (2008). An approach to rationalize partitioning sample size into individual regions in a multi-regional trial. Drug Information Journal 42:139–147.

Ko, F. S., Tsou, H. H., Liu, J. P., Hsiao, C. F. (2010). Sample size determination for a specific region in a multiregional trial. Journal of Biopharmaceutical Statistics 20:870–885.

Ministry of Health, Labor, and Welfare. (2007). Basic Principles on Global Clinical Trials. Tokyo, Japan.

Quan, H., Li, M. Y., Chen, J., Gallo, P., Binkowitz, B., Ibia, E., Tanaka, Y., Ouyang, S. P., Luo, X. L., Li, G., Menjoge, S., Talerico, S., Ikeda, K. (2010). Assessment of Consistency of Treatment Effects in Multiregional Clinical Trials. Drug Information Journal 44:617–632.

Quan, H., Li, M., Shih, W. J., Ouyang, S. P., Chen, J., Zhang, J., Zhao, P. L. (2012). Empirical shrinkage estimator for consistency assessment of treatment effects in multi-regional clinical trials. Statistics in Medicine 32:1691–1706.

Quan, H., Zhao, P. L., Zhang, J., Roessner, M., Aizawa, K. (2010). Sample size considerations for Japanese patients in a multi-regional trial based on MHLW guidance. Pharmaceutical Statistics 9:100–112.

Teng, Z., Chang, M. (2016). Optimal multiregional clinical trials. Chapter 11. Multi-regional clinical trials for simultaneous global new drug development. CRC Press.

Teng, Z., Chang, M. (2016). Unified additional requirement in consideration of regional approval for multi-regional clinical trials. Accepted by Journal of Biopharmaceutical Statistics.

Tsou, H. H., Hung, H. M. J, Chen, Y. M., Huang, W. S., Chang, W. J., Hsiao, C. F. (2012). Establishing consistency across all regions in a multi-regional clinical trial. Pharmaceutical Statistics 11:295–299.

Tsong, Y., Chang, W. J., Dong, X. Y., Tsou, H. H. (2012). Assessment of regional treatment effect in a multiregional clinical trial. Journal of Biopharmaceutical Statistics 22:1019–1036.

A Statistical Decision Framework Applicable to Multipopulation Tailoring Trials

Brian A. Millen

Abstract Interest in tailored therapeutics has led to innovations in trial design and analysis. Trials now range from traditional overall population trials with exploratory subgroup analysis to single population tailoring trials, to multipopulation tailoring trials. This paper presents an overview of the trial options and provides a framework for decision making in confirmatory multipopulation tailoring trials.

Keywords Type I error rate • Influence condition • Interaction condition • Personalized medicine

1 Introduction

It is becoming increasingly recognized in the medical community that populations of patients with particular illnesses are made up of multiple, sometimes latent, subpopulations in which response to treatment differs. As a logical result, interest in tailored therapeutics—having treatments available to patients according to subgroup status—has grown greatly. These subpopulations may be defined by clinical characteristics, gene or protein expression, demographics, or other markers. Clinical trials and their associated analyses have adapted according to this interest in developing tailored therapies. As depicted in Fig. 1 below, trials now range from traditional overall population trials with exploratory subgroup analyses, to single population tailoring trials (i.e., trials in a defined subpopulation of patients), to multipopulation trials (i.e., trials studying an overall population of patients which provides inference for both the overall population and a predefined subgroup(s) of patients).

A prominent example of a single population tailoring trial is found in the confirmatory trastuzumab Herceptin® trials for breast cancer. In those trials, only patients whose tumors demonstrated high levels of HER-2 expression were enrolled. The combined results of these trials indicated a highly significant benefit for

B.A. Millen (✉)
Eli Lilly and Company, Lilly Corp. Center; Indianapolis, IN, USA
e-mail: bmillen@lilly.com

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_10

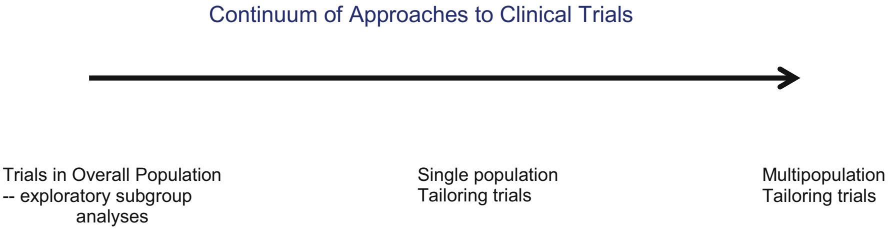

Continuum of Approaches to Clinical Trials
Trials in Overall Population -- exploratory subgroup analyses
Single population Tailoring trials
Multipopulation Tailoring trials

Fig. 1 Continuum of approaches for studying subpopulations in clinical trials

Herceptin in the studied patient population (Romond et al. 2005). Of course, these trials are unable to address the question if a broader population of breast cancer patients would benefit from Herceptin treatment.

While single population tailoring trials offer efficient study in targeted subpopulations, they fail to provide contrasting information in the complementary (marker negative) subpopulation. This information is generally desirable. Nonetheless, single population tailoring trials are appropriate in some contexts, considering weight of evidence for a predictive marker and weighing risk and benefit to patients.

A clear alternative to simple overall population trials or single population tailoring trials are multipopulation tailoring trials. These trials enroll patients in the overall population, yet are designed to enable inference for both the overall population and a predefined subpopulation(s). (For simplicity throughout, we focus on the case of a single subpopulation of interest.) A prominent example is the SATURN trial of erlotinib (Cappuzzo et al. 2010). This trial was designed to evaluate the effect of erlotinib as maintenance therapy in the population of patients with unresectable non-small cell lung cancer as well as in a subpopulation of patients whose tumors overexpressed epidermal growth factor receptor. The trial was positive in both the overall population and the predefined subpopulation. The resulting US approved label (Tarceva US product label 2013) for erlotinib Tarceva® includes the results for the overall population, as well as for the predefined subpopulation and the complementary subpopulation. It is an *enhanced label*, per the nomenclature of Millen et al. (2012, 2014a).

Given the efficiency and richness of inference of multipopulation tailoring trials, they present an option to be considered in any clinical development plan wherein there is potential evidence of a predictive marker of efficacy.

Of course, the multiple inferential outcomes of multipopulation trials require additional design and analysis considerations compared with the single population trials. These considerations are all related to decision error rate control. This includes standard familywise error rate control associated with multiple testing, as well as control of two potential errors associated with multipopulation trials: influence error rate and interaction error rate. These are briefly addressed in the next section.

2 Decision Principles

In addition to the standard consideration of type I error control, Millen et al. (2012) introduced two analysis principles to protect against decision errors in multipopulation tailoring trials. These are referred to as the *influence condition* and the *interaction condition*, providing protection against influence and interaction decision errors, respectively.

The influence condition states that to support a claim of broad (overall population) effect, there must be evidence that the beneficial effect of treatment is not limited to only the predefined subpopulation. The application of the influence condition helps minimize the likelihood of influence errors (i.e., concluding an overall population effect when the effect is solely driven by the subpopulation).

The interaction condition states that to achieve simultaneous claims in the overall population as well as the predefined subpopulation, there must be evidence of a differential effect between the predefined subpopulation and the complementary subpopulation. Else claims of effect would be limited to the broad, overall population. Application of the interaction condition helps minimize the likelihood of an interaction error (i.e., concluding that a relevant differential effect between subpopulations which does not exist).

An algorithm for decision-making based on hypothesis testing plus these conditions is given in Millen et al. (2014a) and reproduced in the diagram and text below. Mathematical formulations for evaluating the influence and interaction conditions appear in Millen et al. (2014a, b).

As depicted in the diagram, an algorithm for decision-making may proceed sequentially as follows:

Step 1: Conduct the primary hypothesis tests for the overall population and predefined subpopulation according to an appropriate multiple testing procedure which provides strong control of the type I error rate.

Step 2: Assuming the primary hypothesis test for effect in the overall population is positive, then assess the influence condition. If the influence condition is not satisfied, then conclude the positive treatment effect is limited to the predefined subpopulation and recommend a subpopulation-only indication. If the influence condition is satisfied, then go to Step 3.

Step 3: Assess the interaction condition. If the interaction condition is not satisfied, then recommend a claim of primary effect only at the broad population level. If the interaction condition is satisfied, then recommend an enhanced label which supports treatment for the broad population and provides information on the differential subpopulation effects.

Details of all possible outcomes are noted in the decision diagram (Fig. 2).

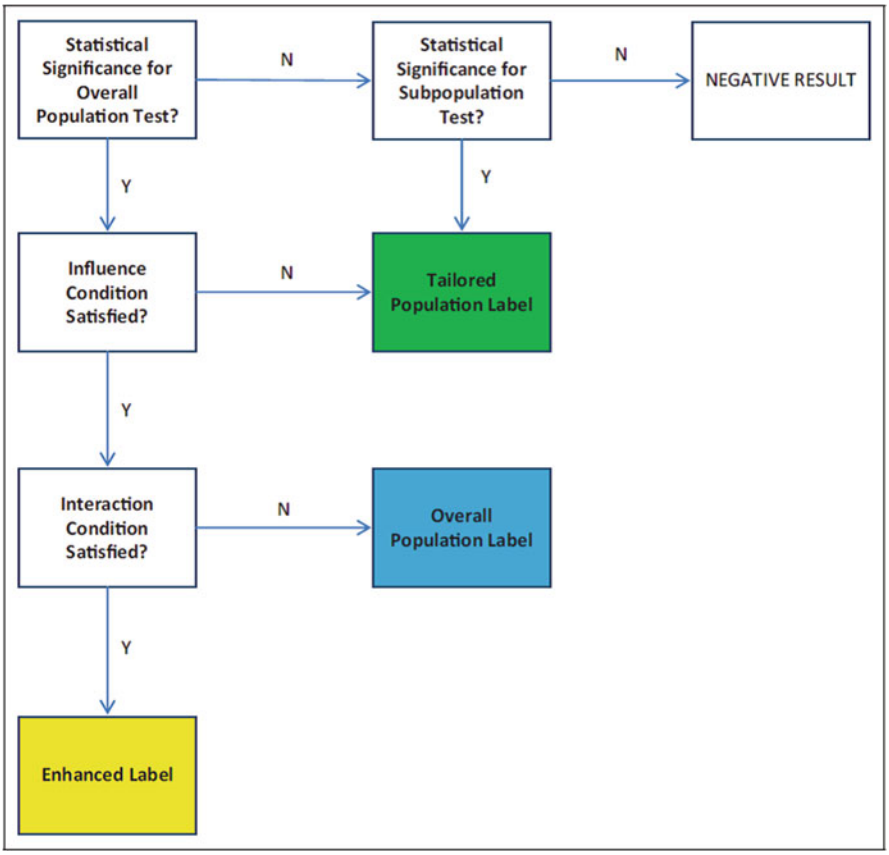

Statistical Significance for Overall Population Test?
N
Statistical Significance for Subpopulation Test?
N
NEGATIVE RESULT
Y
Y
Influence Condition Satisfied?
N
Tailored Population Label
Y
Interaction Condition Satisfied?
N
Overall Population Label
Y
Enhanced Label

Fig. 2 Statistical decision algorithm applicable to confirmatory multipopulation tailoring trials

3 Discussion and Summary

In this paper we presented a framework for inference in multipopulation tailoring trials. While the decision principles and framework are readily implemented for any multipopulation trial, analysis details will differ for the various endpoint types. Ding et al. (2016) address the notion of subgroup mixable estimation. The crux of their work is that the overall population parameter should be expressable as a mixture of the subpopulation parameters in order for inference from such trials to satisfy reasonable logical expectations. For example, if the primary endpoint is a mean difference, it easily satisfies the mixability criterion. For time to event analyses, they recommend use of ratio of median survival times, rather than hazard ratios, as the former satisfies this condition while the latter does not. This underscores the careful selection of analysis endpoints when considering multipopulation trials. Similarly, at the design stage of multipopulation trials, the researcher will want to ensure appropriate sample sizes at the subpopulation level to enable acceptable likelihoods of meeting the multiple inferential objectives outlined, including satisfying the influence and interaction conditions.

The discussion in this paper has focused on a trial-level decision paradigm for multipopulation tailoring trials. While indication and labeling recommendations in some cases are based on single trial paradigms, they are more often based on *programs* consisting of two or more pivotal trials. The decision principles outlined in this paper remain applicable to the program setting. Particular guidance for program-level application is the subject of future work.

In conclusion, the use of multipopulation tailoring trials offer an important option for delivering on the promise of tailored therapeutics. The decision framework and concepts outlined in this paper support the trialist's consideration and implementation of such trials.

References

Romond EH, Perez EA, Bryant J, et al. Trastuzumab plus adjuvant chemotherapy for operable HER-2 positive breast cancer. *New England Journal of Medicine.* 2005; 353: 1673–1684.

Cappuzzo F, Ciuleanu T, Stelmakh L, et al. Erlotinib as maintenance treatment in advanced non-small-cell lung cancer: a multicentre, randomized, placebo-controlled phase 3 study. *Lancet Oncology.* 2010; 11:521–529.

[Tarceva US product label]. http://www.accessdata.fda.gov/drugsatfda_docs/label/2010/021743sl4s16lbl.pdf. Accessed April 6, 2013.

Millen BA, Dmitrienko A, Mandrekar SJ, Zhang Z, and Williams DA. Multipopulation Tailoring Clinical Trials: Design, Analysis, and Inference Considerations. *Therapeutic Innovation and Regulatory Science.* 2014; 48: 453–462.

Millen BA, Dmitrienko A, Ruberg S, and Shen L. A statistical framework for decision making in confirmatory multipopulation tailoring trials. *Therapeutic Innovation and Regulatory Science.* 2012; 46: 647–656.

Millen BA, Dmitrienko A, and Song G. Bayesian assessment of the influence and interaction conditions in multipopulation tailoring clinical trials. *Journal of Biopharmaceutical Statistics*. 2014; 24: 94–109.

Ding Y, Lin H-M, and Hsu JC. Subgroup mixable inference on treatment efficacy in mixture populations, with an application to time-to-event outcomes. *Statistics in Medicine*. 2016; 35: 1580–1594.

Assessing Benefit and Consistency of Treatment Effect Under a Discrete Random Effects Model in Multiregional Clinical Trials

Jung-Tzu Liu*, Chi-Tian Chen*, K.K. Gordon Lan*, Chyng-Shyan Tzeng, Chin-Fu Hsiao, and Hsiao-Hui Tsou

Abstract The traditionally uniform treatment effect assumption may be inappropriate in an multiregional clinical trial (MRCT) because of the impact on the drug effect due to regional differences. Lan and and Pinheiro (2012) proposed a discrete random effects model (DREM) to account the treatment effects heterogeneity among regions. However, the benefit of the overall drug effect and the consistency of the treatment effect in each region are two major issues in MRCTs. In this article, the

*Author contributed equally with all other contributors.

J.-T. Liu
Institute of Population Health Sciences, National Health Research Institutes, 35 Keyan Road, Zhunan, Miaoli County 350, Taiwan

Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu, Taiwan

C.-T. Chen
Institute of Population Health Sciences, National Health Research Institutes, 35 Keyan Road, Zhunan, Miaoli County 350, Taiwan

K.K.G. Lan
Janssen Pharmaceutical Companies of Johnson & Johnson, Raritan, NJ, USA

C.-S. Tzeng
Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu, Taiwan

C.-F. Hsiao (✉)
Institute of Population Health Sciences, National Health Research Institutes, 35 Keyan Road, Zhunan, Miaoli County 350, Taiwan
e-mail: chinfu@nhri.org.tw

H.-H. Tsou (✉)
Institute of Population Health Sciences, National Health Research Institutes, 35 Keyan Road, Zhunan, Miaoli County 350, Taiwan

Graduate Institute of Biostatistics, College of Public Health, China Medical University, Taichung, Taiwan
e-mail: tsouhh@nhri.org.tw

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_11

power of benefit is derived under DREM and the overall sample size determination in an MRCT. Comparison of DREM and traditional continuous random effects model (CREM) is also illustrated here. In order to assess the treatment benefit and consistency simultaneously under DREM, we consider the concept of the Method 2 in "Basic Principles on Global Clinical Trials" guidance to construct the probability function of benefit and consistency. We also optimize the sample size allocation to reach maximum power for the benefit and consistency.

Keywords Multiregional clinical trial • Discrete random effects model • Consistency • Power for benefit and consistency • Optimal sample size allocation

1 Introduction

Developing pharmaceutical products via multiregional clinical trials (MRCTs) has become a standard strategy. Traditionally, the treatment effect was assumed to have a fixed positive value over all regions in an MRCT (Kawai et al. 2008; Ko et al. 2010; Tsou et al. 2010, 2011, 2012). However, regional heterogeneity in MRCTs has been observed and may have an impact upon a medicine's treatment effect (Hung et al. 2010; Wang and Hung, 2012). To account for heterogeneous treatment effect across regions in an MRCT, the random effects model with a normal prior, originally proposed by DerSimonian and Laird in the context of meta-analysis (DerSimonian and Laird, 1986), has been used by many authors to combine treatment effect estimates (Chen et al. 2012; Quan et al. 2010). Here, we denote "random effects model (REM) with a normal prior" as "CREM," with the capital C indicating the "Continuous prior." CREM usually combines and weights the regional treatment effects by the inverse of the within- region variance to obtain the global estimate. Hung et al. (2010) used CREM to explore the impact of regional differences on the efficiency of trial design and found the problem on design insufficiency and sample size implication as a result of ignoring regional differences when in truth there are regional differences. Recognizing that regional treatment differences are typically not random samples from a normal distribution, Lan and Pinheiro (2012) proposed a discrete random effects model (DREM) to account for regional heterogeneity. The proposed DREM may be more applicable in certain practical situations and provide a compromise between the fixed effects model and CREM.

2 Discrete Random Effects Model for Heterogeneous Treatment Effect

We consider the cases of two parallel treatment arms: a test product T and a placebo control C, with a 1:1 patient allocation ratio. Larger response indicates better outcome. Assume that the patient population is partitioned into s disjoint clinical regions S_1, S_2, $\ldots$, S_s, with $P(S_i) = W_i$ and $\sum_i W_i = 1$. For $j = T$, C, the sample size N_j is the sum of the s regional sample sizes, $N_j = \sum_{i=1}^{s} N_{ij}$. We further assume the total sample size in a treatment group is $N_T = N_C = N$, and that $N_{ij} = N_i$ for $j = T$, C. Theoretically, the sample sizes from the s regions are random; but in practice they can be replaced by the observed values.

The DREM is described in the following. For the kth patient in treatment j allocated to Region i, the treatment effect for a randomly selected patient in the population is μ_{ijk} and the weight $P(\mu_{ijk} = \delta_{ij}) = W_i$, for $i = 1, 2, \ldots, s$. The distribution F_j for μ_{ijk} is then defined by the possible values $\{\delta_{1j}, \delta_{2j}, \ldots, \delta_{sj}\}$ and their respective probabilities $\{W_1, W_2, \ldots, W_s\}$, for $j = T$, C. The overall treatment difference, δ $(=\delta_T - \delta_C)$, is defined as the weighted sum of the individual effect of regional difference, denoted by $\delta = \sum_{i=1}^{s} W_i \delta_i$, where $\delta_i = \delta_{iT} - \delta_{iC}$ is the overall average treatment effect in the ith region. Let an individual treatment response X_{jk} for the kth subject in treatment j be expressed as $X_{jk} = \mu_{jk} + \varepsilon_{jk}$, where μ_{jk} and ε_{jk} are assumed to be independent, $k = 1, 2, \ldots, N_j, j = T$, C. Suppose that μ_{jk} follows a discrete distribution F_j, with mean δ_j and variance τ^2. The parameter $\varepsilon_{jk} \sim N(0, {\sigma_j}^2)$. Correspondingly, the overall within-group variance $\sigma^2 = \sum_{i=1}^{s} W_i \sigma_i^2$, where $\sigma_i^2 =$ within group variance in the ith region, and the between-region variance τ^2 due to the discrete prior is $\tau^2 = \sum_{i=1}^{s} W_i (\delta_i - \delta)^2$. When the regional sample sizes N_i are reasonably large, the overall treatment difference δ can be adequately estimated by $\widehat{\delta} = \sum_i w_i \widehat{\delta}_i$. Here w_i $(=N_i/N)$ and $\widehat{\delta}_i$ $(= (\sum_{k=1}^{N_i} X_{iTk} - \sum_{k=1}^{N_i} X_{iCk})/N_i)$ are the estimators of W_i and δ_i, respectively. In other words, the estimated overall treatment effect $\widehat{\delta}$ will depend on the percentage of subjects contributed by different regions. Note that the weights W_i are unknown parameters and that they are not random (Lan et al. 2014). The variance of $\widehat{\delta}$ is estimated by $\text{var}(\widehat{\delta}) = 2(\sigma^2 + \tau^2)/N$. The variances τ^2 and σ^2 can be adequately estimated by $\widehat{\tau}^2 = \sum_i w_i \widehat{\delta}_i^2 - \widehat{\delta}^2$ and $\widehat{\sigma}^2 = \sum_k w_i \widehat{\sigma}_i^2$, respectively, where $\widehat{\sigma}_i^2$ is the sample variance in the ith region.

3 Hypothesis, Power for Benefit, and Sample Size Determination

The hypotheses for testing the overall treatment effect can be written as

$$H_0 : \delta \leq 0 \quad \text{versus} \quad H_A : \delta > 0. \tag{1}$$

Let $\widehat{\delta} = \sum_i w_i \widehat{\delta}_i$. Under DREM, the test statistic is given by

$$Z = \frac{\widehat{\delta}}{\sqrt{\text{var}\left(\widehat{\delta}\right)}} = \frac{\widehat{\delta}}{\sqrt{2\left(\sigma^2 + \tau^2\right)/N}} \tag{2}$$

For the fixed effects model, $\tau^2 = 0$. The distribution of the test statistic Z approximates normal as $N \to \infty$. The null hypothesis H_0 is rejected at the significance level α and the treatment T is claimed beneficial if the test statistic $Z > z_\alpha$, where z_α is a (1−α) quantile of the standard normal distribution. Therefore, the power function for benefit is given by

$$PB = P\,[\text{Benefit}] = P\left(Z > z_\alpha \middle| N, \delta\right) = \Phi\left(\frac{\delta}{\sqrt{2\left(\sigma^2 + \tau^2\right)/N}} - z_\alpha\right) \tag{3}$$

where Φ denotes the cumulative probability function of the standard normal distribution and $z_\alpha \approx 1.96$ if one-sided $\alpha = 0.025$. This equation can be used in a fixed sample design; but the concept can also be extended to group sequential designs.

3.1 Sample Size Determination Under DREM

Under DREM, the total required sample size N is planned for detecting an expected treatment difference $\delta = \delta^* > 0$ at significance level α and power $1-\beta$ satisfying

$$\sqrt{\frac{N}{2\left(\sigma^2 + \tau^2\right)}} = \frac{z_\alpha + z_\beta}{\delta *} \tag{4}$$

Thus, the required sample size can be determined by

$$N_{\text{DREM}} = 2\left(\frac{z_\beta + z_\alpha}{\delta *}\right)^2 \left(\sigma^2 + \tau^2\right) \tag{5}$$

Under the fixed effects model, $\tau^2 = 0$. Thus, the sample size is determined as follows

$$N_{\text{FIX}} = 2\left(\frac{z_\beta + z_\alpha}{\delta *}\right)^2 \cdot \sigma^2.$$

3.2 Sample Size Determination Under CREM

An alternative model considered in many articles to account for the heterogeneous treatment effect across regions is a random effects model with a continuous prior.

The model assumptions are

$$\widehat{\nu}_k \middle| \nu_k \sim N\left(\nu_k, 2\sigma^2/N_{\text{CREM},k}\right) \quad \text{and} \quad \nu_k \sim N\left(\nu, \tau^2\right)$$

where $N_{\text{CREM},k}$ is sample size per arm of a subgroup k. A corresponding variance of $\widehat{\nu}_k$ is estimated by $\text{var}(\widehat{\nu}_k) = 2\sigma^2/N_{\text{CREM},k} + \tau^2$. Similarly, the overall treatment difference ν is a weighted sum of the individual effect of regional difference, given by $\nu = \sum_{k=1}^{K} W_k \nu_k$. The overall treatment difference ν is estimated by $\widehat{\nu} = \sum_{k=1}^{K} w_k^* \widehat{\nu}_k$ with the "reciprocal of the variance" weight $w_k^* = [\text{var}(\widehat{\nu}_k)]^{-1} / \sum_{k=1}^{K} [\text{var}(\widehat{\nu}_k)]^{-1}$. The variance of $\widehat{\nu}$ is estimated by $\text{var}(\widehat{\nu}) = \sum_{k=1}^{K} \left(w_k^*\right)^2 \text{var}(\widehat{\nu}_k) = 1/\sum_{k=1}^{K} [\text{var}(\widehat{\nu}_k)]^{-1}$. Under CREM, the required sample size N_{CREM} for each arm under CREM can be derived by following equation,

$$\left(\frac{\delta}{z_{1-\beta} + z_{1-\alpha/2}}\right)^2 = \frac{1}{\sum_{k=1}^{K} \left(\frac{2\sigma^2}{p_k \, N_{\text{CREM}}} + \tau^2\right)^{-1}},$$

where $p_k = N_{\text{CREM},k}/N_{\text{CREM}}$ is sample-size proportion for subgroup k and $\sum_{k=1}^{K} p_k = 1$.

3.3 Comparison Between DREM and CREM

To explore the impact of regional difference in designing an MRCT, the sample size determination can be evaluated by the sample-size ratios for CREM and DREM, respectively, as $N_{\text{CREM}}/N_{\text{FIX}}$ and $N_{\text{DREM}}/N_{\text{FIX}}$. We use a numerical example to illustrate sample size comparisons for three models. Assume that three regions are entertained in an MRCT and given $p_1 = p_2 = 0.25$, $p_3 = 0.5$, $\sigma = 25$, $\alpha = 0.05$, and $1-\beta = 0.8$.

Figure 1 displays the sample-size ratio as a function of effect size δ/σ for DREM and CREM compared with the fixed effects model, and two types of ratio τ/σ (=0.2 and 0.5). It is clear that the sample size decreases with increasing value of δ/σ and increases with increasing value of τ/σ for each model. When the between-region variance is small relative to the within-region variance (say, $\tau/\sigma = 0.2$), the sample-size ratio $N_{\text{DREM}}/N_{\text{FIX}} = 1.04$ for any values of δ/σ. On the contrary, the ratio $N_{\text{CREM}}/N_{\text{FIX}} > 1.7$ when $\tau/\sigma = 0.2$ and $\delta/\sigma \leq 0.5$. When the between-region variance is large relative to the within-region variance (say, $\tau/\sigma = 0.5$), the sample-size ratio $N_{\text{DREM}}/N_{\text{FIX}} = 1.25$ for any values of δ/σ. On the contrary, the ratio $N_{\text{CREM}}/N_{\text{FIX}} > 2.4$ when $\tau/\sigma = 0.5$ and $\delta/\sigma \leq 1$.

Figure 2 shows the sample-size ratio as a function of effect size δ/σ for different models. Given fixed value δ, the ratio $N_{\text{CREM}}/N_{\text{FIX}}$ increases much faster than the ratio $N_{\text{DREM}}/N_{\text{FIX}}$, while the values of between-region variance τ are increased. As seen in Fig. 2, the sample-size ratio $N_{\text{DREM}}/N_{\text{FIX}}$ is always close to 1.

For $0 \leq \tau/\sigma \leq 1.2$, and given $\delta = 20$, $\sigma = 25$, $p_1 = p_2 = 0.25$, $p_3 = 0.5$, $\sigma = 25$, $\alpha = 0.05$, and $1-\beta = 0.8$, plot of sample-size ratio for CREM ($N_{\text{CREM}}/N_{\text{FIX}}$) and DREM ($N_{\text{DREM}}/N_{\text{FIX}}$) against values of τ/σ are given in Fig. 3. The sample size of DREM (N_{DREM}) is larger than N_{FIX} when the variance component τ is large

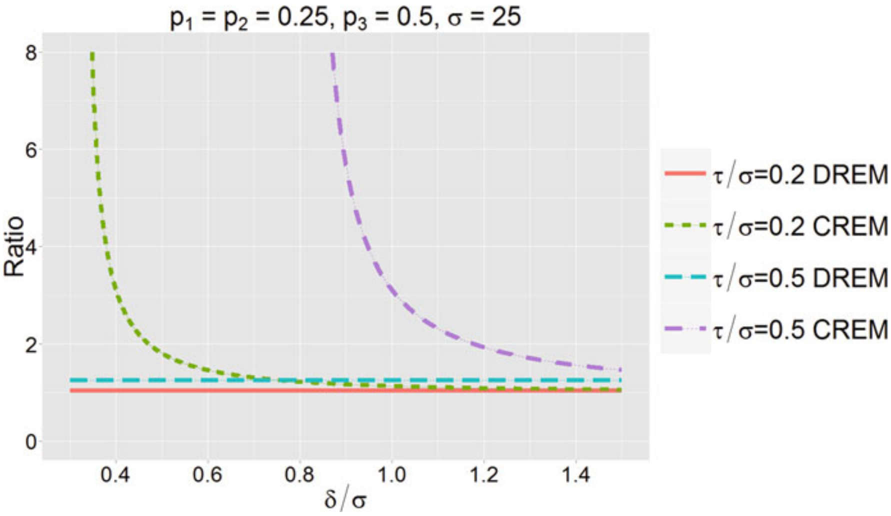


Fig. 1 Sample-size ratio $N_{\text{DREM}}/N_{\text{FIX}}$ and $N_{\text{CREM}}/N_{\text{FIX}}$ versus δ/σ at $\tau/\sigma = 0.2$ and 0.5

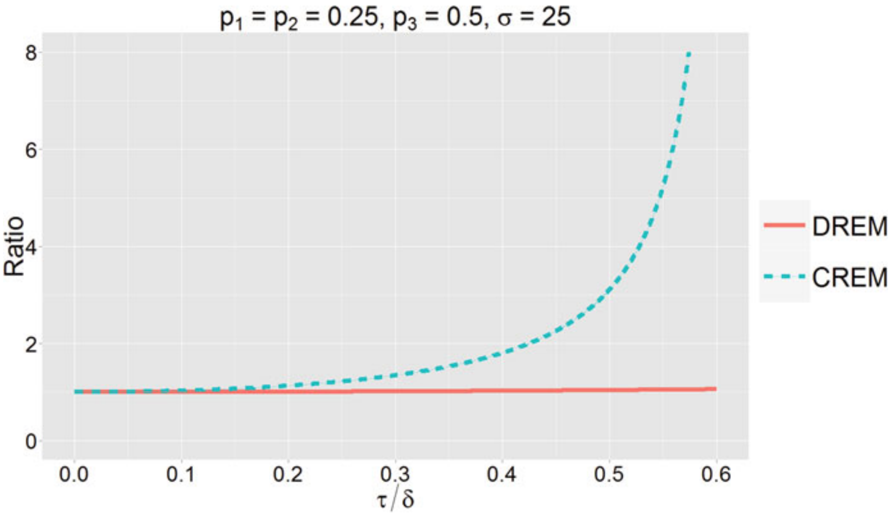


Fig. 2 Sample-size ratio $N_{\text{DREM}}/N_{\text{FIX}}$ and $N_{\text{CREM}}/N_{\text{FIX}}$ versus τ/δ

relative to σ. For example, the ratio of DREM reaches twofold when $\tau/\sigma = 1$. In Fig. 3, $N_{\text{CREM}}/N_{\text{FIX}}$ increases severely compare to $N_{\text{DREM}}/N_{\text{FIX}}$ as the between-region variance τ increases. For example, for $\tau/\sigma = 0.4$, the ratio of CREM reaches 3.2-fold compared with the fixed effects model; but the ratio of DREM reaches only 1.2-fold. Hence, heterogeneity among regions needs to be considered in establishing an MRCT.

As explored in the numerical example, the sample size estimated by the conventional fixed effects model (N_{FIX}) may be insufficient for an MRCT when

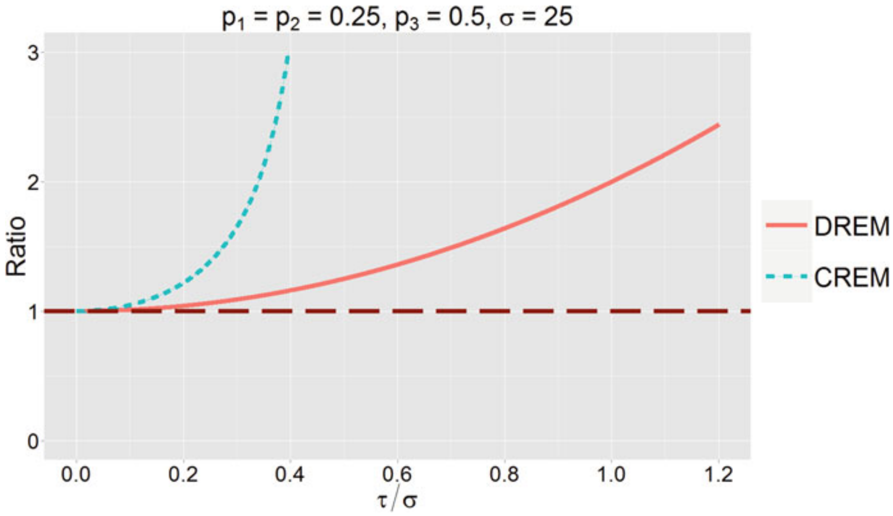


Fig. 3 Sample-size ratio $N_{\text{DREM}}/N_{\text{FIX}}$ and $N_{\text{CREM}}/N_{\text{FIX}}$ versus τ/σ

the between region variance τ is relatively large. The sample size N_{DREM} estimated under DREM may be more applicable for detecting the overall treatment effect $\delta > 0$.

4 Probability for Benefit and Consistency

For the planning and implementation of global clinical studies, the Japanese Ministry of Health, Labour and Welfare (MHLW) published a guideline "Basic Principles on Global Clinical Trials" in 2007 (Ministry of Health & Labour and Welfare of Japan (MHLW), 2007). The guideline includes a method for assessing the consistency among all participating regions in an MRCT (Method 2). That is, (M2): $\widehat{\delta}_i > 0$, for all $i = 1, 2, \ldots, s$. We now consider power for benefit and consistency based on Method 2 (M2). Let us define the following notations.

$$PC = P\left[\text{Consistency (M2)} \middle| N, \theta\right] = P\left(\widehat{\delta}_i > 0, \text{ for all } i \middle| N, \theta\right),$$

$$PBC = P\left[\text{Benefit \& Consistency (M2)} \middle| N, \theta\right]$$
$$= P\left(\text{Z} > 1.96 \,\&\widehat{\delta}_i > 0, \text{ for all } i \middle| N, \theta\right),$$

where $\theta = (\sigma^2, \tau^2, \delta_1, \ldots, \delta_S, W_1, \ldots, W_s)$, W_i denotes the proportion of patients, $i = 1, 2, \ldots, s$, and $\sum W_i = 1$.

4.1 Probability for Consistency (PC) Under DREM

Kawai et al. (2008) derived the probability of observing positive regional effects for all regions (Consistency (M2)) under a fixed effects model as follows.

$$PC_{\text{Fix}} = P\left[\widehat{\delta}_1 > 0, \ldots, \widehat{\delta}_s > 0\right\} = \prod_{i=1}^{s} \Phi\left(\sqrt{W_i} \cdot \left(z_\alpha + z_\beta\right)\right) \tag{6}$$

As seen in Eq. 6, PC_{Fix} depends only on the number of regions s and the proportion of patients W_i in different regions. We derived the power for consistency (PC) based on M2 of MHLW under DREM as follows. First, define the test statistic $Z_i = \frac{\widehat{\delta}_i}{\sqrt{2\left(\sigma^2+\tau^2\right)/NW_i}}$, for $i = 1, \ldots, s$, and define $Z_{\min} = \min\{Z_1, Z_2, \ldots, Z_s\}$. Then $Z = \sum_{i=1}^{s} \sqrt{W_i} Z_i$ and the power for consistency (PC) under DREM is given by

$$PC \ = P\left[\text{Consistency (M2)} \,\middle|\, N, \theta\right] = P\left(\text{Z}_{\min} > 0\right) \tag{7}$$

Using Eq. 4, we can write Eq. 7 as

$$PC = \prod_{i=1}^{s} \Phi\left(\frac{\sqrt{W_i} \cdot \delta_i}{\delta} \cdot \left(z_\alpha + z_\beta\right)\right). \tag{8}$$

As seen in Eq. 7 and Eq. 8, PC under DREM depends on the number of regions, sample size proportions $\{W_i\}$, the treatment effects $\{\delta_i\}$ in all regions, within-group variance σ^2, and between-region variance τ^2. Moreover, PC is an increasing function of the effect size $\delta_i/(\sigma^2 + \tau^2)$ for a fixed N.

4.2 Optimal Allocation Among Regions to Maximize PC

In planning an MRCT, researchers need to determine weights W_i and gather the observed information from each region for demonstrating the efficacy and consistency of a drug. We provide an optimal allocation of patients among regions to maximize the power of consistency (PC) under DREM. We found that PC under DREM is maximized when $\sqrt{W_1} \cdot \delta_1 = \cdots = \sqrt{W_s} \cdot \delta_s$ (Liu et al. 2016). Thus, the optimal allocation $\{W_i\}$ depends only on the values of $\{\sqrt{W_i} \cdot \delta_i\}$ under DREM. Special case: If all regional effects are equal, then PC is maximized when $W_1 = \ldots = W_s = 1/s$. This result is reduced to the finding in Kawai et al. (2008).

4.3 Probability for Benefit and Consistency (PBC)

Now, we define a power for benefit and consistency (PBC) to describe the regional power with a consistent trend (M2) of MHLW 2007. For a fixed N, the power for benefit and consistency (PBC) is denoted by $PBC(N) = \mathrm{P}[Z > z_\alpha \ \& \ Z_{\min} > 0 \mid N]$. Let parameter space $\Theta = \{\text{All } \theta = (W_i, \delta_i, i = 1, 2, \ldots, s) \text{ under consideration}\}$. We found some interesting observations as follows.

(1) PBC increases with N for any fixed θ (Liu et al. 2016).
(2) $PC(N, \theta)$ increases as W_1 increases.
(3) $PBC(N, \theta)$ increase as W_1 increases for given $PB(N) = 80\,\%$.

Moreover, we provide three algorithms for deriving sample size at the desired level of power for benefit and consistency (Liu et al. 2016).

5 Summary

The heterogeneous regional treatment effects among regions were observed from many MRCTs. In this article, we have introduced an alternative random effects model, the discrete random effects model (DREM) to account for heterogeneous treatment effect across regions. We have also addressed consideration of a consistent trend across regions under DREM after showing the overall efficacy of a drug in all global regions. We provide an approach to have enough patients in each region so that the chance of observing a positive treatment in each region reaches a desired level, say 80 %. In practice, regional treatment effects are unknown. Our approach could also provide some guidelines on the design of MRCTs with consistency when the regional treatment effect are assumed to fall into a specified interval. Our approach could serve as a starting point to discuss the scientific rationale for deciding the number of subjects for different regions in a multiregional trial.

Acknowledgements This work was supported by grants from the National Health Research Institutes, Zhunan, Taiwan (PH-102-PP-11, PH-103-PP-11, PH-104-PP-11).

References

Kawai N, Stein C, Komiyama O, Li Y. An approach to rationalize partitioning sample size into individual regions in a multiregional trial. *Drug Information Journal* 2008; **42**:139–147.

Ko FS, Tsou HH, Liu JP, Hsiao CF. Sample size determination for a specific region in a multi-regional trial. *Journal of Biopharmaceutical Statistics* 2010; **24**: 870–885.

Tsou HH, Chow SC, Lan KKG, Liu JP, Wang M, Chen HD, Ho LT, Hsiung CA, Hsiao CF. Proposals of statistical consideration to evaluation of results for a specific region in multi-regional trials — Asian Perspective. *Pharmaceutical Statistics* 2010; **9**:201–206.

Tsou HH, Chien TY, Liu JP, Hsiao CF. A consistency approach to evaluation of bridging studies and multiregional trials. *Statistics in Medicine* 2011; **30**:2171–2186.

Tsou HH, Hung HMJ, Chen YM, Huang WS, Chang WJ, Hsiao CF. Establishing consistency across all regions in a multi-regional clinical trial. *Pharmaceutical Statistics* 2012; **11**: 295–299.

Hung HMJ, Wang SJ, O'Neill RT. Consideration of regional difference in design and analysis of multi-regional trials. *Pharmaceutical Statistics* 2010; **24**:173–178.

Wang SJ, Hung HMJ. Ethnic sensitive or molecular sensitive beyond all regions being equal in multiregional clinical trials. *Journal of Biopharmaceutical Statistics* 2012; **22**:879–893.

DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**: 177–88.

Chen CT, Hung HMJ, Hsiao CF. Design and evaluation of multiregional trials with heterogeneous treatment effect across regions. *Journal of Biopharmaceutical Statistics* 2012; **22**:1037–1050.

Quan H, Zhao PL, Zhang J, Roessner M, Aizawa K. Sample size considerations for Japanese patients in a multi-regional trial based on MHLW Guidance. *Pharmaceutical Statistics* 2010; **9**:100–112.

Lan KKG, Pinheiro J. Combined estimation of treatment effects under a discrete random effects model. *Statistics in Biosciences* 2012; **4**:235–244.

Lan KKG, Pinheiro, J, Chen F. Designing multiregional trials under the discrete random effects model. *Journal of Biopharmaceutical Statistics* 2014; **24**:415–428.

Ministry of Health, Labour and Welfare of Japan (MHLW). Basic principles on global clinical trials 2007. (Available at: http://www.pmda.go.jp/kijunsakusei/file/guideline/new_drug/GlobalClinicalTrials_en.pdf) [Accessed date: May 6, 2014].

Liu JT, Tsou HH, Lan KK Gordon, Chen CT, Lai YH, Chang WJ, Tzeng CS, Hsiao CF. Assessing the Consistency of the Treatment Effect under the Discrete Random Effects Model in Multiregional Clinical Trials. *Statistics in Medicine* 2016 (In press).

Design and Analysis of Multiregional Clinical Trials in Evaluation of Medical Devices: A Two-Component Bayesian Approach for Targeted Decision Making

Yunling Xu, Nelson Lu, and Ying Yang

Abstract Current statistical design and analysis of multiregional clinical trials for medical devices generally follows a paradigm where the treatment effect of interest is assumed consistent among US and OUS regions. In this paper, we discuss the situations where the treatment effect might vary among US and OUS regions, and propose a two-component Bayesian approach for targeted decision making. In this approach, anticipated treatment difference among US and OUS regions is formally taken into account by design, hopefully leading to increased transparency and predictability of targeted decision making.

Keywords Medical devices • Multiregional clinical trials • Two-component Bayesian approach • Regional law

1 Introduction

More and more sponsors are using clinical data collected from studies conducted both in the United States (US) and outside the US (OUS) to support premarket approval (PMA) for medical devices (Lu et al. 2011). OUS data are often collected as part of multiregional clinical trials (MRCTs), i.e. clinical trials conducted simultaneously in multiple geographical or regulatory regions under the same clinical protocol. Both sponsors and the US regulators are embracing the idea of utilizing MRCT data to speed up the process for US patients to access the effective and safe medical devices. Utilization of MRCTs also stimulates the convergence of quality standard across regions. In doing so, a variety of challenges may arise, such as statistical issues for design, conduct, monitoring and analysis of MRCTs, especially when there are substantial regional differences in treatment effects.

Y. Xu (✉) • N. Lu • Y. Yang
Division of Biostatistics, Center for Devices and Radiological Health, Food and Drug Administration, 10903 New Hampshire Ave., Silver Spring, MD 20993, USA
e-mail: yun-ling.xu@fda.hhs.gov

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_12

In this paper, we propose a two-component Bayesian approach for targeted decision making utilizing MRCT data. The paper is organized as following. In Sect. 2, two motivation examples are presented to illustrate regional differences in treatment effect. Current statistical practice is discussed and the unmet need for PMA in medical devices is pointed out. In Sect. 3, we propose a two-component Bayesian approach for targeted decision making. In Sect. 4, the proposed two-component Bayesian approach is applied to an example for designing and analyzing MRCTs for targeted decision making. The paper is concluded with Sect. 5 of some remarks.

Please note that we approach the problem from a perspective of regulatory science bounded by US medical device law, rather than a perspective of science without judicial boundary. Some of the considerations could be different between from the two perspectives.

2 Motivation Examples: Current Practice and Unmet Need

Currently, regarding approval of a medical device in the US based on MRCT results, decision making is usually based on the statistical inference on the global treatment effect estimated from the pooled data. Data from US and OUS are pooled together for analysis through pre-specified hypothesis testing for treatment effect. If the global hypothesis test is significant, statistical test of a treatment by region interaction is often utilized to assess treatment effect difference across regions. When there is no evidence showing regional difference in treatment effect, the decision making is mostly straightforward regardless of the significance of the global hypothesis test. However, when there is evidence showing regional differences in treatment effects, the current practice in statistical design and analysis of MRCTs has its limitation in facilitating decision making. Let's look at two examples where the current practice of statistical analysis works less effectively to quantify the uncertainty of a regulatory decision. These examples are constructed for illustrating problems frequently encountered in our review experience.

Example 1 An MRCT was conducted to evaluate the effectiveness of a cardiac resynchronization therapy with biventricular pacing for symptomatic patients. The primary endpoint is the proportion of subjects with improved clinical symptoms within a year after implantation of the investigational device. The control group patients were treated by cardiac resynchronization therapy with right ventricular pacing. The global analysis indicates that 85.6 % of patients with biventricular pacing improve their clinical symptoms as opposed to only 78.1 % of the control patients improve their clinical symptoms. The p-value of t-test is 0.015, showing the superiority of the biventricular pacing. The subsequent test for treatment by region interaction yields a p-value of 0.13 showing evidence of heterogeneous treatment effects between US and OUS. The descriptive results (Table 1) stratified by US and OUS indicates that, although the investigational device seems to work OUS, it is

Table 1 Proportion of subjects with improved clinical symptoms by treatment group and region

US		OUS	
Treatment	Control	Treatment	Control
80 % (100/125)	79 % (46/58)	88 % (250/284)	78 % (111/143)

Table 2 Quality of life (QOL) score change from baseline by treatment group and region

	Average ± SD (N)	
Region	Treatment	Control
US	1.85 ± 1.17 (59)	1.21 ± 1.20 (29)
Region A	1.18 ± 1.15 (76)	0.67 ± 1.21 (40)
Region B	1.38 ± 1.13 (55)	1.78 ± 1.36 (29)

unclear if the same can be said in the US. One notable factor is that the patients are managed differently in medical practice between US and OUS as the investigational device had been previously approved and used in the OUS region.

Example 2 An MRCT was conducted to investigate an investigational device for treating depression comparing to a control (standard of care). The primary endpoint is a quality of life questionnaire (QOL) score on a 7-point scale where a higher score represents a better quality of life. For the pre-specified primary effectiveness analysis of the QOL score change from baseline based on global analysis, the p-value of t-test for superiority of treatment (1.44) to control (1.16) is 0.035, which is slightly larger than the pre-specified significance level of 0.025. Although the global analysis failed to show the superiority, an interaction test was conducted, and the treatment by region interaction had a p-value$<$0.01. Post hoc analyses by region results are presented in Table 2. Substantial improvement (p-value$=$0.01) is observed for patients in the US regarding QOL score for the investigational device group compared to the control, while patients in region B do not show improvement. All sites followed the same protocol with uniform inclusion and exclusion criteria, and no plausible reasons have been found to account for the notable regional difference in treatment effects. In summary, global test for superiority has failed due to regional difference in treatment effect although the investigational device seems to work in the US. Thus the statistical results are less supportive of approval under current practice.

In example 1, an important quantity that would greatly inform the decision making is the probability of the product being falsely approved for the US population based on the global analysis. In example 2, an important quantity that would greatly inform the decision making is the probability of the product being falsely not approved for the US population based on the global analysis. For OUS counter parts, symmetric questions could be asked for. These quantities cannot be easily estimated under current practice with the working assumption that the treatment effects are the same across regions. This is because, under this assumption, the operating characteristics of the design, including false approvable and approvable rates, are the same across different regions. Therefore, there is a need to develop a statistical framework for designing (and analyzing) MRCTs with the following features: (1) formal consideration of anticipated and unanticipated treatment effect difference across regions; and (2) evaluable regional specific operating characteristics for varying degrees of regional difference in treatment effect.

When regional difference in treatment effect exists and is taken into account, the operating characteristics of the design become different among regions, i.e. each region has its own different operating characteristics, including false approvable and approvable rates. Furthermore, in practice the regulations guiding medical device approval are very different among regions. For example, for the approval of a Class III medical device in the US, reasonable assurance of safety and effectiveness must be demonstrated as indicated by section 513(a)(1)(C) of the Federal Food, Drug, and Cosmetic Act (FD & C Act 2015); in EU, demonstration of the device effectiveness is not required for CE (Conformité Européenne or Communauté Européenne) marking (CE Mark 2015). That is, evidence is usually only required on device functional performance (i.e. that the device functions as intended) as opposed to clinical performance (that patients benefit from the treatment) (Kramer et al. 2012). To consider both the potential regional difference in treatment effect and the existing regional difference in statutory requirements in regulatory approval, we propose the concept of *targeted decision making* in using MRCT data for approval of medical devices. Targeted decision making means here a decision based on region specific operating characteristics such as false approvable/approvable rates which are evaluated by taking into account regional differences in treatment effect and regional differences in statutory requirement. In the next section, a two-component Bayesian approach is proposed for targeted decision making.

3 A Two-Component Bayesian Approach for Targeted Decision Making

For convenience of description, let i index region and δ_i (bigger is better) be the treatment effect of the investigational device against control in region i (=US, OUS). The discussion here focuses on the issue of US vs. OUS as an example of targeted US decision making. On one hand, as there are shared commonalities between the US and the OUS populations, OUS data are informative about δ_{us}; on the other hand, there could be differences between the US and the OUS populations, so OUS data may be less informative about δ_{us} than US data. As such, two tiers of evidence can be defined: direct evidence and supporting evidence. Direct evidence is designated here as evidence from data collected in the US where the study population matches the intended population as characterized by its unique intrinsic and extrinsic factors; supporting evidence is designated here as evidence from data from all regions where study population may have different intrinsic and extrinsic factors from the intended population in the US. The direct evidence and supporting evidence can be statistically quantified as a direct evidence measure and a supporting evidence measure, respectively.

In this paper, we use Bayesian posterior probability to serve as direct evidence measure and supporting evidence measure: Direct evidence measure, denoted by Pd, is the posterior probability of ($\delta us > 0$) given US data; supporting evidence measure, denoted by Ps, is the posterior probability of ($\delta us > 0$) given US and OUS data. Utilizing these two evidence measures, a two-component statistical decision framework can be formulated for targeted US decision making: (1) Direct Evidence Criterion (DEC) is defined as Pd > t1 and Ps > t2, and (2) Supporting Evidence Criterion (SEC) is defined as Pd > t3 and Ps > t4. Notations t1, t2, t3 and t4 are threshold values decided within a specific regulatory decision context at the design stage (usually t1 $\geq$ t3 and t2 $\leq$ t4). A targeted decision making can be supported by meeting either the DEC or the SEC.

Regarding the threshold parameters in DEC, t1 is devised to define the required significant direct evidence; similarly in SEC, t4 is devised to define the required significant supporting evidence. When direct evidence is significant in DEC, adequate supporting evidence is required as such t2 is devised; similarly when supporting evidence is significant in SEC, adequate direct evidence is required as such t3 is devised. Therefore, t1 should be set greater than or equal to t2, and t3 should be set no more than t4. Please note that the two-component framework implicitly contains a minimum requirement for direct evidence and for targeted consistency in treatment effect across regions; and a balance between the two minimum requirements. For medical device studies, t1, t2, t3 and t4 need to be pre-specified at study design stage based on clinical and regulatory considerations within a specific context. The threshold values of t1, t2, t3 and t4 essentially determine the extent and the way of utilizing data outside of the region for the targeted decision making. If more direct evidence is required to support a regulatory decision making, t1 and t3 should be set at greater values. Correspondingly, if a regulatory decision making can be supported by mostly supporting evidence, the threshold values t1 and t3 could be set lower as needed. In general, the larger the anticipated regional treatment effect difference, the more direct evidence needed. Furthermore, the size of a region could limit the amount of direct evidence required for a targeted decision making. For example, a region with small sample size may have to mostly rely on SEC with lower threshold values of t3 and higher threshold value of t4 for regulatory decision making.

With the two-component Bayesian approach for targeted US regulatory decision making utilizing MRCT, false approval (rejecting $H_0^{US}|H_0^{US}, H^{OUS}$) rate and false disapproval (failing to reject $H_0^{US}|H_a^{US}, H^{OUS}$) rate are of the direct interest. Here H^{OUS} represent all possible different treatment effects across all OUS regions. A thorough exploring over all concerned scenarios is necessary to help understand the impact of anticipated and unexpected regional differences on the targeted US regulatory decision making. In doing so, all the "what if ..." questions from all the stakeholders shall be addressed, and thus the probability of anticipated regret could be controlled at a desirable level through an appropriate study design. The threshold values of t1, t2, t3 and t4 should be chosen through simulations to accommodate all concerned scenarios and control the probabilities of anticipated regret.

4 Operating Characteristics Driven Study Design Process with the Two-Component Bayesian Approach

The proposed two-component Bayesian approach for targeted decision making is a tool for designing MRCTs, but it is not a design per se. It is an operating characteristics-driven iterative study design process requiring intensive simulations. During the design process, all concerned possible scenarios regarding true state of heterogeneity (δ_{US}, δ_{OUS}) from all involved stakeholders should be considered. For each scenario (combination of δ_{US} and δ_{OUS}), the operating characteristics, including false approvable and approvable probabilities for the targeted regulatory decision making, shall be calculated and evaluated. If the stakeholders are not satisfied with the false approvable and approvable probabilities with respect to concerned scenarios, sample size and/or threshold values can be adjusted and the associated operating characteristics will then be reevaluated. This process shall be repeated until the operating characteristics are desirable for all stakeholders regarding all concerned scenarios.

We here illustrate the study design process with a hypothetical example. Suppose that a two-arm, randomized controlled superiority MRCT is planned to be conducted in 4 regions (US, R2, R3 and R4) with a randomization ratio of 1:1 within each region. The clinical response endpoint follows a normal distribution. For convenience, let δk be the true treatment effect in region k (k = 1 (US), 2(R2), 3(R3), 4(R4)).

Pd is calculated from the following model as P(β <0|data from region 1):

Let i index subject in group t (treatment)/c (control) from region 1

$\{$
$y_{ci} \sim N\ (\mu_1{}^c,\ \tau),\quad y_{ti} \sim N\ (\mu_1{}^t,\ \tau),$
$\mu_1{}^t = \mu_1{}^c + \beta$
$\tau \sim \text{Gamma}\ (0.001,\ 0.001)$
$\mu_1{}^c,\ \beta \sim N\ (0,\ 1000)$
$\}$

Ps is calculated from the following Bayesian random effect model as P(β <0|data from all regions):

Let i index subject in group t (treatment)/c (control) from region k

$\{$
$y_{cki} \sim N\ (\mu_k{}^c,\ \tau),\quad y_{tki} \sim N\ (\mu_k{}^t,\ \tau)$
$\mu_k{}^t = \mu_k{}^c + \beta_k$
$\mu_k{}^c \sim N\ (\mu_c,\ \tau_c),\quad \beta_k \sim N\ (\beta,\ \tau_t)$
$\tau,\ \tau_c,\ \tau_t \sim \text{Gamma}\ (0.001,\ 0.001)$
$\mu_c,\ \beta \sim N\ (0,\ 1000)$
$\}$

Simulations were performed in R. BRugs, an interface to the OpenBUGS, to calculate the posterior probabilities with MCMC sampling (BRugs 2015). For the control group, the clinical response follows a normal distribution with mean of 2 and precision of 0.01. Each trial was repeated 10,000 times to get operating characteristics.

4.1 Calibrated Target Setting

We use current practice as the starting point for sample size planning: assuming that the true treatment effect is δ ($\mu^t - \mu^c = -2$) for all regions, 400 subjects per arm are needed to have a power of 80 % with one-sided α of 0.025, using a two-sample *t*-test.

With the sample size of 400, suppose that after a thorough clinical consideration of potential factors that might cause regional difference in treatment effect, the US regulatory agency calls for at least half of the sample size being from the US. With additional considerations regarding OUS regions, the sample size is allocated according to a 5(US):2(R2):2(R3):1(R4) ratio.

First, extensive trial and error simulations are performed to decide a set of threshold values for DEC and SEC with the goal being about 80 % US approvable rate when each of the 4 regions has the same treatment effect of −2 shown as Design A in Table 3. For Design A, the false approvable rate in the US is about 5.37 % when every region has a true treatment effect of 0 and threshold values of t's are as the following: DEC (t1 = 0.975, t2 = 0.85) and SEC (t3 = 0.85, t4 = 0.975). Then, extensive simulations were performed to decide a set of threshold values for DEC and SEC with the goal being about 2.5 % approvable rate in the US when every region has a true treatment effect of 0 as shown under Design B in Table 3. For Design B, the approvable rate in the US is about 61.9 % when every region has a true treatment effect of −2 and threshold values of t's are as the following: DEC (t1 = 0.975, t2 = 0.85) and SEC (t3 = 0.95, t4 = 0.975). Design C shown in Table 3 has a similar design as Design B but with an increase of 120 subjects in US. With the extra sample size, the false approval rate in the US becomes 2.96 % when each of the 4 regions has a true treatment effect of 0, while the approval rate in the US is increased to 79.6 % when every region has a true treatment effect of –2.

4.2 Extended Evaluation

At the calibrated target setting stage, the designs are evaluated for the US false approvable and approvable rates under scenarios that the treatment effects are the same across all regions. Suppose that Designs A and C in Table 3 are decided for extended evaluation. At this extended evaluation stage, we compare the US approval rate under various scenarios of heterogeneous treatment effects across regions.

Table 3 US False approvable and approvable rates for three candidate study designs: simulation results

	DEC and SEC threshold					Approvable rate (%) in US under	
Design	t1	t2	t3	t4	Sample size (US:R2:R3:R4)	$\mu^t - \mu^c = 0$ for all regions	$\mu^t - \mu^c = -2$ for all regions
A	0.95	0.85	0.85	0.95	400 (5:2:2:1)	5.37	78.8
B	0.975	0.85	0.95	0.975	400 (5:2:2:1)	2.45	61.9
C	0.975	0.85	0.95	0.975	520 (8:2:2:1)	2.96	79.6

Table 4 Comparison of Design A and C in terms of US false approval rates with varying magnitude of treatment effect in other regions: simulation results

Treatment effect ($\mu^t - \mu^c$)				US approval rate (%)			
	OUS			Design A		Design C	
US	R2	R3	R4	Two-component Bayesian	Pooled t-test	Two-component Bayesian	Pooled t-test
0	0	0	0	5.37	2.75	2.96	2.59
0	0	−2	−2	9.22	13.08	3.98	5.58
0	−2	−2	−2	11.25	28.54	4.72	23.40

Design A: DEC (t1 = 0.95, t2 = 0.85), SEC (t3 = 0.85, t4 = 0.95), Sample size 400 per arm with allocation ratio 5:2:2:1 (US:R2:R3:R4). Design C: DEC (t1 = 0.975, t2 = 0.85), SEC (t3 = 0.95, t4 = 0.975), Sample size 520 per arm with allocation ratio 8:2:2:1 (US:R2:R3:R4). Pooled t-test: one-sided significance level of 0.025

Table 5 Comparison of Design A and C in terms of US approval rates with varying reduced magnitude of treatment effect in other regions: simulation results

Treatment effect ($\mu^t - \mu^c$)				US approval rate (%)			
	OUS			Design A		Design C	
US	R2	R3	R4	Two-component Bayesian	Pooled t-test	Two-component Bayesian	Pooled t-test
−2	−2	−2	−2	78.8	80.1	79.6	89.5
−2	−1	−1	−1	72.9	56.0	77.2	73.1
−2	−2	0	0	71.7	50.8	76.9	69.4
−2	0	0	0	65.3	29.2	74.1	50.8

Design A: DEC (t1 = 0.95, t2 = 0.85), SEC (t3 = 0.85, t4 = 0.95), Sample size 400 per arm with allocation ratio 5:2:2:1 (US:R2:R3:R4). Design C: DEC (t1 = 0.975, t2 = 0.85), SEC (t3 = 0.95, t4 = 0.975), Sample size 520 per arm with allocation ratio 8:2:2:1 (US:R2:R3:R4). Pooled t-test: one-sided significance level of 0.025

Table 4 shows the operating characteristics of Design A and C under the scenarios of no treatment effect in the US and various magnitudes of treatment effects in other regions. It can be observed that, when there is no treatment effect in all regions, US false approvable rate is higher for Design A comparing to Design C. With varying magnitude of treatment effect in OUS regions, US false approval rate for both Design A and C becomes higher; but Design C does so in a lower speed comparing to Design A. For both designs, US false approval rate with the two-component Bayesian approach is lower than with the pooled t-test as the US treatment effect becomes distinctly different from the OUS regions.

Table 5 shows the operating characteristics of Design A and Design C under the scenarios of treatment effect in the US being −2 and various reduced magnitude of treatment effects in other regions. It can be observed that, with same treatment effects across all regions, US approvable rate is similar between two designs. For other cases, US approvable rates under both designs become lower; and such rates under Design A are lower than the associated rate under Design C. For both designs, US approvable rates using the two-component Bayesian approach are much higher than associated rates using the pooled t-test when the US treatment effect becomes distinctly different from the OUS regions.

Summarizing the results regarding false approvable rate and approvable rate in the US from Tables 4 and 5, it can be seen that (1) currently used pooled t-test analysis performs better if true treatment effects are the same across regions; (2) the proposed two-component Bayesian approach performs better with heterogeneous treatment effects across regions; and (3) Design C performs better than Design A as larger sample size is demanded in Design C (320 vs. 200).

If little regional differences in treatment effects among regions are expected and/or enrollment of the trial is anticipated to be challenging, Design A may be a more logical choice than Design C.

5 Discussion

In this paper, a two-component Bayesian approach is proposed for targeted decision making by taking into account of possible regional differences in treatment effects at the study design stage. It is devised from the perspective of pre-market applications for medical device in the US as treatment effects are often suspected to vary among regions for many types of medical devices. The proposed approach is best suited for cases where heterogeneous treatment effects are anticipated across regions. It also aims to safeguard against the situations where the unexpected regional treatment effect differences would cause great difficulties in the decision making. The idea behind this approach is to use DEC and SEC criteria to set a balance between evidence exhibited from US and OUS data. By doing so, it is implied that a minimal sample size needs to be allocated to the US to ensure the applicability of the study conclusion to the US population. In general, substantial expected regional differences would warrant a higher threshold for direct evidence measure. An explicit decision tree with defined operating characteristics increases the predictability and transparency of targeted decision making.

Consistency assessment among regions is a regulatory consideration in utilizing MRCTs for PMAs. In recent years, various definitions for consistency have been proposed and studied (for a sample of references, see Ministry of Health. Labour and welfare of Japan (MHLW) 2007; Chen et al. 2010; Quan et al. 2010, 2014; Tsou et al. 2012) for MRCTs. Such consistency assessment is usually done after the demonstration of the treatment effect by the pooled analysis, and is usually not incorporated into the overall evaluation of operating characteristics of the MCRT study design for targeted decision making. The proposed approach in this paper implies targeted consistency in treatment effects across regions in the sense that, when the evidence solely based on US data is not strong enough, stronger evidence from OUS is needed; when the evidence from US data is strong, adequate evidence is needed from OUS data. The level of targeted consistency is embed in the DEC and SEC.

The proposed approach in this paper is flexible to accommodate different statutory requirements by different regulatory agencies. The flexibility lies with setting region specific DEC and SEC. Practically, by the size of the target population, a big region can afford relying on more direct evidence if warranted; a small region is more likely to rely more on supporting evidence. In addition, a balance shall be stroked between the amount of the targeted region versus out of the targeted region data within a specific regulatory decision context, which could include the risk of the devices and the anticipated impact of varying intrinsic and extrinsic factors between the targeted region and out of the targeted region on the effectiveness of the investigational device.

For assessing the supporting evidence, in general, a Bayesian random effect model with non-informative prior can be used; and the Bayesian random effect model can incorporate patient /region level covariates if available. At study design stage, it is necessary to pre-define the regions in a MRCT and ideally to have randomization stratified by region to facilitate assessing the strength of supporting evidence. The geographic area under the US FDA jurisdiction would form one region for effectiveness evaluation; OUS regions could be pre-defined by various criteria. One is to be formed according to judicial areas. Another is to be formed across judicial boundaries according to similarity in intrinsic and extrinsic factors, such as medical practice and healthcare policy in particular, as discussed by Binkowitz (2010). It should be noted that there cannot be too few OUS regions as it will be difficult to quantify the variability in treatment effect across regions.

Medical device trials have some features that are different from drug trials. First, there are many types medical products/practice that can serve as the control in a medical device trial. Possible controls include a “no treatment”, a placebo, a medical device, a drug, drug management, and standard of care (Design Considerations for Pivotal Clinical Investigations for Medical Devices 2015). Unlike the standard placebo control in a drug trial, a ‘no treatment’ or placebo control in a device trial may be unfeasible or may be unethical due to unacceptable risk to the patient. Second, in many medical device trials, blinding (masking) is often practically impossible for patients, health care professionals, and/or investigators, and sometimes blinding (masking) could jeopardize subject care. Furthermore, for medical devices, the physician’s skill and the need for multiple specialties can be important treatment effect modifiers. In addition, the regulations governing medical device approval are very different between the US and other regions such as European Union (EU) (Kramer et al. 2012). All these features could cause the heterogeneous treatment effects for medical device across regions (for a more detailed account, see Xu et al. 2016).

Ideally when substantial regional differences are foreseen, MRCTs should not be used. However, in reality, the law requires FDA, in deciding whether to approve, license, or clear a drug or device, to accept data from clinical trials conducted in OUS regions, provided that the applicant successfully demonstrates that the data are adequate under FDA’s approval standards (FADASIA 2014). Therefore, we proposed the two-component Bayesian approach in this paper for situations where regional heterogeneous treatment effects are anticipated, or where it is warranted to safeguard against unexpected regional treatment effect differences. The two-component Bayesian approach is not intended to be scientifically ideal, but pragmatically to help inform a regional regulatory decision making in rather a possible difficult situation, especially where regional heterogeneous treatment effects are due to unknown, or known but not observed factors.

Acknowledgement We would like to thank an anonymous reviewer whose comments helped improving this manuscript.

Disclaimer No official support or endorsement by the Food and Drug Administration of this article is intended or should be inferred.

References

Lu N, Nair R, Xu Y. Decision rules and associated sample size planning for regional approval utilizing multi-regional clinical trials. Presented at: the 4th Annual FDA/MTLI Medical Device and IVD Statistical Issues Workshop, National Harbor, MD, 13-14 April, 2011.

Food, Drug and Cosmetic Act (FD & C Act), www.fda.gov/RegulatoryInformation/Legislation/FederalFoodDrugandCosmeticActFDCAct/FDCActChapterVDrugsandDevices/ucm110188.htm (accessed May 25, 2015).

CE Mark, http://eur-lex.europa.eu/LexUriServ/site/en/consleg/1993/L/01993L0042-20031120-en.pdf (accessed October 22, 2015).

Kramer DB, Xu S, Kesselheim AS. Regulation of Medical Devices in the United States and European Union. *The New England J Medicine* 2012; 366: 848-855.

BRugs: R interface to the OpenBUGS MCMC software. http://cran.r-project.org/web/packages/BRugs/index.html (accessed June 30, 2015).

Ministry of Health. Labour and welfare of Japan (MHLW). Basic Principles on Global Clinical Trials 2007. http://www.pmda.go.jp/english/service/pdf/notifications/0928010_e.pdf (accessed July 8, 2015).

Chen J, Quan H, Binkowitz B, et al. Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review. *Pharmaceutical Statistics* 2010; 9: 242–253.

Quan H, Chen J, Gallo P, et al. Assessment of consistency of treatment effects in multiregional clinical trials. *Drug Information Journal* 2010; 44:617–632.

Quan H, Mao X, Chen J, et al Multi-regional clinical trial design and consistency assessment of treatment effects. *Statistics in Medicine* 2014; 33:2191-2205.

Tsou H-H, Hung HMJ, Chen Y-M, et al. Establishing consistency across all regions in a multi-regional clinical trial. *Pharmaceutical Statistics* 2012; 11:295–299.

Binkowitz B. Highlights from the PhRMA MRCT Key Issue Team & DIA MRCT Workshop. Presented at: the 4th Seattle Symposium in Biostatistics: Clinical Trials. Seattle, WA, 20-23 November, 2010.

Design Considerations for Pivotal Clinical Investigations for Medical Devices http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM373766.pdf (accessed February 26, 2015).

Xu Y, Lu N, Gray G. Special Considerations for Medical Devices: An Overview. Multi-Regional Clinical Trials for Simultaneous Global New Drug Development. Chen J, Quan H. (Eds.), 2016, Chapman and Hall/CRC: 229-309.

FADASIA: http://www.gpo.gov/fdsys/pkg/BILLS-112s3187enr/pdf/BILLS-112s3187enr.pdf (accessed February 3, 2014).

Semiparametric Analysis of Interval-Censored Survival Data with Median Regression Model

Jianchang Lin, Debajyoti Sinha, Stuart Lipsitz, and Adriano Polpo

Abstract Analysis of interval censored survival data has become increasingly popular and important in many areas including clinical trials and biomedical research. Generally, right censored survival data can be seen as a special case of interval censored data. However, due to the fundamentally special and complex nature of interval censoring, most of the commonly used survival analysis methods for right censored data, including methods based on martingale-theory (Andersen et al., Statistical models based on counting processes. Springer, New York, 1992), can not be used for analyzing interval censored survival data. Most of the popular semiparametric models for interval censored survival data focus on modeling the hazard function. In this chapter, we develop a semiparametric model dealing with the median regression function for interval censored survival data, which introduce many practical advantages in real applications. Both semiparametric maximum likelihood estimator (MLE) and the Markov chain Monte Carlo (MCMC) based semiparametric Bayesian estimator, including how to incorporate the historical information, have been proposed and presented. We illustrate the case study through a real breast cancer data example and make a comparison between different models. Key findings and recommendations are also discussed to provide further guidance on application in clinical trials.

J. Lin (✉)
Takeda Pharmaceuticals, Cambridge, MA 02139, USA
e-mail: jianchang.lin@takeda.com

D. Sinha
Department of Statistics, Florida State University, Tallahassee, FL 32306, USA
e-mail: sinhad@stat.fsu.edu

S. Lipsitz
Division of General Medicine, Brigham and Womens Hospital, Boston, MA 02115, USA
e-mail: slipsitz@partners.org

A. Polpo
Department of Statistics, Federal University of São Carlos, São Carlos, SP 13565-905, Brazil
e-mail: polpo@ufscar.br

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_13

Keywords Survival analysis • Interval-censored • Quantile regression • Semiparametric analysis

1 Introduction

Interval censored survival data has become increasingly common problems in many areas including financial, epidemiological, medical, and sociological research studies. A typical example of interval censored data is medical or epidemiological studies of slow-growth diseases that have no immediate outward symptoms (Sun 2006). Usually, the occurrence of interval censored observation is mainly due to the nature of the disease and/or the structure of the study design. For such studies (for example, Finkelstein and Wolfe 1985), the survival time T_i of patient i can not be observed, but can only be determined to be within an interval $(A_i, B_i]$ of a sequence of clinic visit or examination times. Two special cases of interval-censored data are found in current status data (Jewell and van der laan, and others), where either $A_i = 0$ or $B_i = +\infty$. In recent years, there has been a lot of interest and research activity for the models and appropriate analysis for such data. For a more comprehensive review, we refer to the authoritative book by Sun (2006) and the references therein.

Generally, right censored survival data can be seen as a special case of interval censored data, and some of the inference approaches based on right censored data can be applied to interval censored data directly or with some minor modification. However, due to the fundamentally special and complex nature of interval censoring, most of the commonly used survival analysis methods, including methods based on martingale-theory (Andersen et al. 1992), can not be used for analyzing interval censored survival data. Most of the popular semiparametric models for interval censored survival data focus on modeling the hazard function $h(t \mid x_i)$ given covariate x_i (For example, Sun 2006; Finkelstein and Wolfe 1986; Satten et al. 1998; Pan 2000; So et al. 2010). For semiparametric Bayesian analysis, there are existing works including Sinha et al. (1999) and Ghosh and Sinha (2000) dealing with Cox's model, and Hanson and Johnson (2004) and Hanson and Yang (2007) using accelerated failure time model. However, for studies with interval-censored data, focusing on the effects of covariate x_i either on instantaneous risk or change of time-scale may not be appropriate because the design of such studies does not allow continuous monitoring of survival. Also, the Bayesian semiparametric procedures require the knowledge of the prior mean function of the baseline function, either hazard or survival. Often in practice, the available prior information about the study under consideration is only related to certain quantiles, for example the median, of the survival response. This prior information about the median and anticipated range of the change of median for different values of covariate x_i are routinely elicited and then used for power and sample size evaluations of the study (Piantadosi 2005).

In this chapter, we focus on the inference and theoretical properties of a semiparametric regression survival model with a transform $g_\lambda(\cdot)$ on both the $\log(T_i)$ and regression function $\eta_i = \beta x_i$. We specify a symmetric unimodal nonparametric error distribution for the transformed response $g_\lambda(\log(T))$. This leads to a semiparametric model with log-linear regression function $\exp(\beta x_i)$ for the median of T_i. We develop our inference procedure and associated theoretical justifications within the semiparametric likelihood as well as full Bayes paradigms. To our knowledge there is no previous works dealing with the median regression function for interval censored survival data. Commonly used self-consistency based and martingale based estimating equations for median regression (Cheng et al. 1997; Yang and Prentice 1999; McKeague et al. 2001; Bang and Tsiatis 2003; Portnoy 2003; Peng and Huang 2008), have not been extended to deal with interval-censoring. Our computational algorithms for the semiparametric maximum likelihood estimator and the Markov chain Monte Carlo (MCMC) based semiparametric Bayesian estimator are easy to implement. Section 2 describes the transformation both-side model and its properties in the context of interval censored data. The semiparametric method is developed in Sect. 3. In Sect. 4, we present an example and make a comparison between different models. We discuss our key findings and comments in Sect. 5.

2 Semiparametric Models

For interval censored data, the observations are from subject $i = 1, \ldots, n$ with $\{T_i, x_i\}$, where T_i is random variable representing the failure time of a subject in the study, $x_i = (1, x_{i1}, \ldots, x_{ir})^T$ is the corresponding vector of r covariates along with the intercept term. For interval censored variable T_i, only an interval $(A_i, B_i]$ is observed such that

$$T_i \in (A_i, B_i], \tag{1}$$

where $A_i \leq B_i$. In the following chapter, $A_i = B_i$ represents a exact observation and $B_i = \infty$ means a right censored observation. We assume that the interval censoring mechanism is "non-informative", which means the process $\{N_{0_i}(t)\}$ of observation times, clinic visits, and T_i are independent given x_i.

Bickel and Doksum (1981) define a monotone power transformation family, an extension of the Box-Cox power family (Box and Cox 1964), as

$$g_\lambda(y) = \frac{\text{sign}(y)|y|^\lambda}{\lambda}, \tag{2}$$

where $\lambda > 0$ and $\text{sign}(y) = 1$ if $y \geq 0$ and $\text{sign}(y) = -1$ if $y < 0$. Our transform both-side model assumes that $g_\lambda(\log(T_i))$, for an optimal λ, is symmetric unimodal with median $g_\lambda(\eta_i)$, that is,

$$g_\lambda(\log(T_i)) = g_\lambda(\eta_i) + \varepsilon_i = g_\lambda(\beta x_i) + \varepsilon_i, \tag{3}$$

where, ε_i are independent with unimodal and symmetric around zero distribution F_ε. Carroll and Ruppert (1984), Fitzmaurice et al. (2007), and Lin et al. (2012, 2013) used a parametric transformation both-side regression model for an uncensored continuous response with the original Box-Cox transformation (Box and Cox 1964) and parametric normal density for ε_i.

Since $g_\lambda(\cdot)$ is monotonic increasing with inverse $g_\lambda^{-1}(y) = \text{sign}(y)|y\lambda|^{1/\lambda}$, then

$$P[T_i > \exp(\eta_i)] = P[g_\lambda(\log(T_i)) > g_\lambda(\eta_i)] = P[\varepsilon_i > 0] = 1/2 \tag{4}$$

as long as ε_i has median 0. As a consequence, the survival time T_i has log-linear median $Q_{0\cdot 5} = \exp(\eta_i)$, and any α-percentile of T with $P[T < Q_\alpha] = \alpha$ has the expression

$$Q_\alpha = \exp\{g_\lambda^{-1}(g_\lambda(\eta) + \epsilon_{(\alpha)})\}, \tag{5}$$

where $\epsilon_{(\alpha)}$ is the α-percentile of error distribution F_ε with $P[\varepsilon_i < \epsilon_{(\alpha)}] = \alpha$. In some situations, when we are interested in modeling another percentile Q_{α^*} for $\alpha^* \neq 0 \cdot 5$, instead of median $Q_{0\cdot 5}$, as a log-linear function $\exp\{\eta(x)\} = \exp(\beta x)$ we can modify (3) to assume that unimodal symmetric F_ε satisfies $F_\varepsilon(0) = P[\varepsilon < 0] = \alpha^*$. The expression for other percentiles in (5) does not change, except the function $\exp\{\eta(x)\}$ now corresponds to the α^*-percentile of T. For the rest of the article, we deal with only the median functional for the model in (3). In the last section, we discuss in more detail the differences and advantages of our model compared to existing survival models, including the quantile regression models of Portnoy (2003), Neocleous et al. (2006), and others.

Theorem 1. *Under non-informative interval-censoring, the parameters (λ, β) of the model (3) is identifiable even when F_ε is unknown.*

This important result about identifiability is true even for the restricted case of current status data, when either $A_i = 0$ or $B_i = +\infty$. The proof is based on the fact that $S(t \mid x = 0)$ and $S(t \mid x = 1)$ are identifiable from current status data. Medians of $S(t \mid x = 0)$ and $S(t \mid x = 1)$ identify the parameters (λ, β). We would also like to mention that, for model (3), there exists a unique $(\lambda, \beta, F_\varepsilon)$ satisfying (3). The proof is omitted.

Also, the sign and magnitude of any component of β determines the relationship between every percentile of T and that component of covariate.

Proposition 1. $Q_{\alpha_1}(x) \geq Q_{\alpha_1}(x^*) \Leftrightarrow Q_{\alpha_2}(x) \geq Q_{\alpha_2}(x^*)$, *for all* (α_1, α_2).

The ordering of the percentiles for any two subjects remain same for all α. Without loss of generality, to show this, we take a covariate vector $x = (x_1, x_2)$ with two scalar components. For (3), $\beta_1 > 0 \Rightarrow Q_\alpha(x_1, x_2) > Q_\alpha(x_1^*, x_2)$ for all α as long as $x_1 > x_1^*$ where $Q_\alpha(x_1, x_2)$ is the α-percentile of the survival time for subject with covariate value $x = (x_1, x_2)$. This model property is similar to the uniform ordering of the survival functions property of the Cox model.

3 Model Estimation and Inference

For the model in (3), the likelihood contribution of each subject is $P(T_i \in (A_i, B_i]) = P(\varepsilon \in (\tilde{A}_i, \tilde{B}_i])$, where $\tilde{A}_i = \mathrm{g}_\lambda(\log(A_i)) - \mathrm{g}_\lambda(\eta_i)$ and $\tilde{B}_i = \mathrm{g}_\lambda(\log(B_i)) - \mathrm{g}_\lambda(\eta_i)$. The likelihood based on the observed interval censored data $\mathscr{D}$ is now

$$L(\lambda, \beta, F_\varepsilon \mid \mathscr{D}) = \prod_{i=1}^{n} P(\varepsilon_i \in (\tilde{A}_i, \tilde{B}_i]) = \prod_{i=1}^{n} \left[F_\varepsilon(\tilde{B}_i) - F_\varepsilon(\tilde{A}_i)\right], \tag{6}$$

where $\mathscr{D} = \{T_i \in (A_i, B_i], x_i : i = 1, \ldots, n\}$. An appropriate parametric distribution F_ε, for example, a normal with mean 0 and variance σ^2, results in a parametric likelihood for (6). A numerical method, such as Nelder and Mead (1965) can be used to obtain the maximum likelihood estimator $(\hat{\lambda}, \hat{\beta}, \hat{\sigma})$ of (λ, β, σ), and the corresponding consistent variance of estimate as the inverse of the observed Fisher information matrix. For the parametric Bayesian analysis, the posterior density is $p(\lambda, \beta, \sigma \mid \mathscr{D}) \propto L(\lambda, \beta, \sigma \mid \mathscr{D})\pi(\lambda, \beta, \sigma)$, where $\pi(\lambda, \beta, \sigma)$ is the joint prior density.

The parametric assumption about F_ε may be inappropriate and restrictive in practice. The class of all unknown unimodal symmetric distribution F_ε can be expressed as a scale-mixture of uniforms

$$F_\varepsilon(\epsilon) = \int_0^\infty F_U(\epsilon \mid \theta) \mathrm{d}G(\theta), \tag{7}$$

for some mixing distribution $G(\theta)$, where $F_U(u \mid \theta)$ is a uniform distribution with support $(-\theta, +\theta)$ for $\theta > 0$ (Feller 1971). The full semiparametric likelihood of (λ, β, G) can be derived as

$$L(\lambda, \beta, G \mid \mathscr{D}) \propto \prod_{i=1}^{n} \left[\int_0^{+\infty} \int_{\tilde{A}_i}^{\tilde{B}_i} f_U(\epsilon \mid \theta) \, \mathrm{d}\epsilon \, \mathrm{d}G(\theta) \right]$$

from (6) and (7). For the semiparametric-likelihood analysis, we use an "empirical" version of the above likelihood, where $F_\varepsilon(\epsilon)$ in (7) is replaced with

$$F_\varepsilon(\epsilon) = \sum_{j=1}^{K} p_j F_U(\epsilon \mid \theta_j), \tag{8}$$

where the mixing distribution $G(\theta)$ is discrete with finite support $\Theta = \{\theta_1, \ldots, \theta_K\}$, with unknown $0 < \theta_1 < \cdots < \theta_K$. Maximizing this likelihood of $(\lambda, \beta, \Theta, p)$ is tantamount to maximizing the following likelihood with $K \leq n$.

Theorem 2. *For discrete $G(\cdot)$ with support Θ and* $\mathrm{pr}(\theta = \theta_j) = p_j$ *for* $0 \leq p_j \leq 1$ *and* $\sum_{j=1}^{K} p_j = 1$, *the log-likelihood of (6) is*

$$l(\lambda, \beta, F_\varepsilon \mid \mathscr{D}) = l(\lambda, \beta, p, \Theta) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{K} p_j \frac{\tilde{B}_{ij} - \tilde{A}_{ij}}{2\theta_j} \right] \qquad (9)$$

where $\tilde{A}_{ij} = \min\{\max\{-\theta_j, \tilde{A}_i\}, \theta_j\}$, $\tilde{B}_{ij} = \max\{\min\{\theta_j, \tilde{B}_i\}, -\theta_j\}$.

See Appendix for the proof. Using results of Wong and Severini (1991), under similar regularity conditions, the maximum likelihood estimator $\hat{\beta}$ of the regression parameter has $n^{1/2}$ convergence rate and is asymptotically efficient. This is particularly true because $F_\varepsilon(\epsilon)$ defined in (8) has a density function. The maxima of the likelihood in (6) always exists because this likelihood function is a product of probabilities, a bounded function with $0 \leq L(\lambda, \beta, F_\varepsilon \mid \mathscr{D}) \leq 1$.

Computing the maximum likelihood estimator of $(\lambda, \beta, \Theta, p)$ via directly maximizing (9) may be computationally intensive. To reduce the computational burden, we propose the use of following iterative procedure with two steps in each iteration. At each iteration, we begin with the most recent value of $(\hat{\lambda}, \hat{\beta})$ and $(\widehat{\Theta}, \hat{p})$. We suggest using the parametric maximum likelihood estimator as the initial value of $(\hat{\lambda}, \hat{\beta})$ for the first iteration.

Step 1: Maximize (9) with respect to (Θ, p) to obtain the current value of $(\widehat{\Theta}, \hat{p})$, where $(\lambda, \beta) = (\hat{\lambda}, \hat{\beta})$ is fixed.

At this step, the optimal Θ is the set of ordered distinct values of $\{|\tilde{A}_i|, |\tilde{B}_i| : i = 1, \ldots, n\}$, where $\tilde{A}_i$ and $\tilde{B}_i$ are known functions of $(\hat{\lambda}, \hat{\beta})$, x_i and (A_i, B_i). This implies that the only unknown parameter at this stage is the vector p. There is unique maxima of $l(\hat{\lambda}, \hat{\beta}, p, \Theta \mid \mathscr{D})$ because it is a concave function of p.

Step 2: Considering the values of $(\widehat{\Theta}, \hat{p})$ obtained in Step 1 as fixed, maximize the likelihood in (9) as a function of (λ, β) to obtain a new $(\hat{\lambda}, \hat{\beta})$. Go back to step 1 and continue the iterations until convergence.

This can be implemented using nonlinear optimization algorithms such as Nelder and Mead (1965). This iterative procedure, under mild conditions, maximizes the profile likelihood (Murphy and van der Vaart 2001)

$$l_P(\lambda, \beta \mid \mathscr{D}) = \max_{(p, \Theta)} l(\lambda, \beta, p, \Theta \mid \mathscr{D}),$$

even though $l_P(\lambda, \beta \mid \mathscr{D})$ does not have any closed form expression in this case. The (j, k) component of the estimated limiting covariance matrix of $\hat{\beta}$ is given as

$$-\mathscr{E}_n^2 \Big[l_P(\hat{\lambda}, \hat{\beta} + \mathscr{E}_n u_j + \mathscr{E}_n u_k) - l_P(\hat{\lambda}, \hat{\beta} + \mathscr{E}_n u_j - \mathscr{E}_n u_k)$$
$$-l_P(\hat{\lambda}, \hat{\beta} - \mathscr{E}_n u_j + \mathscr{E}_n u_k) + l_P(\hat{\lambda}, \hat{\beta}) \Big],$$

where u_j is the standard basis vector with 1 in component j, and $\mathscr{E}_n$ is a scalar of order $n^{-1/2}$ (Murphy and van der Vaart 2001).

Standard errors for the maximum likelihood estimate $\hat{\beta}$ are obtained using the inverse of the observed information matrix obtained via numerical differentiation of the likelihood in (9). When the sample size is large, the implementation of step-1 of the algorithm may turn out to be computationally difficult. In this case, we recommend using a penalty function of smoothness on p.

Another option for semiparametric analysis is to use a full Bayes procedure based on Markov chain Monte Carlo samples from the joint posterior

$$\text{pr}(\lambda, \beta, F_\varepsilon \mid \mathscr{D}) \propto L(\lambda, \beta, F_\varepsilon \mid \mathscr{D})\pi_1(\lambda, \beta)\pi_2(F_\varepsilon),$$

where $\pi_1(\lambda, \beta)$ and $\pi_2(F_\varepsilon)$ are the independent, a reasonable assumption of convenience, priors for the parametric part (λ, β) and the nonparametric F_ε, respectively. This full Bayes semiparametric model uses the scale-mixture of uniforms in (7) as the class of symmetric unimodal F_ε, with no requirement on the unknown mixing distribution G being discrete. We use a Dirichlet process (Ferguson 1973) prior $G \sim \text{DP}(G_0, \alpha)$ for unknown $G(\theta)$, where $G_0(\cdot)$ is the known prior mean of $G(\cdot)$ and $\alpha > 0$ is the precision around mean G_0. The following theorem gives us a method of choosing $G_0(\cdot)$ based on the desired form of prior mean $F_0(\cdot)$ of $F_\varepsilon(\cdot)$. Typically, $F_0(\cdot)$ is assumed to be a known, specified a priori, parametric distribution function with corresponding density $f_0(\cdot)$.

Theorem 3. *When* $\text{E}_{prior}[F_\varepsilon(\cdot)] = F_0(\cdot)$, *for* $F_0(\cdot)$ *symmetric around zero with corresponding density* $f_0(\cdot)$, *the corresponding prior mean* $G_0(u)$ *of* $G(\cdot)$ *in (7) is* $\text{d}G_0(u) = -2u\text{d}f_0(u)$, $u > 0$.

The proof follows from a result by Khintchine (1938). To ensure the prior mean of F_ε as $N(0, v^2)$, the functional form of $G_0(u)$ has to be Gamma$(3/2, 1/(2v^2))$. For the implementation of the Markov chain Monte Carlo tool, we use a finite approximation of the constructive definition of the Dirichlet process mixture prior process (Sethuraman 1994) as $F_\varepsilon(\epsilon) \cong \sum_{j=1}^{K} F_U(\epsilon \mid \theta)p_j$, where $\theta_j \sim G_0(\cdot)$, $p_j = V_j \prod_{\ell<j}(1 - V_\ell)$, $V_j \sim \text{Beta}(1, \alpha)$. We use the WinBUGS software (Lunn et al. 2000) to obtain the Markov chain Monte Carlo samples of the posterior distribution $p(\lambda, \beta, F_\varepsilon \mid \mathscr{D})$, and implement a semi-parametric Bayesian analysis in Sect. 4.

4 Data Example

We illustrate our methods via reanalysis of a retrospective study on early stage breast cancer patients (Table 1), who had been treated at the Joint Center for Radiation Therapy in Boston between 1976 and 1980. The data are reported in Finkelstein and Wolfe (1985) with two treatment arms: $x_{1i} = 0$ if patient i received RT (radiation therapy) and $x_{1i} = 1$ if she received RT+CH (radiation therapy + adjuvant

Table 1 Observed intervals in months for times to breast retraction of early breast cancer patients (Sun 2006)

Observed intervals in months
RT (radiation therapy):
(45,], (25,37], (37,], (4,11], (17,25], (6,10], (46,], (0,5], (33,], (15,],
(0,7], (26,40], (18,], (46,], (19,26], (46,], (46,], (24,], (11,15], (11,18]
(46,], (27,34], (36,], (37,], (22,], (7,16], (36,44], (5,12], (38,], (34,]
(17,], (46,], (19,35], (46,], (5,12], (9,14], (36,48], (17,25], (36,], (46,]
(37,44], (37,], (24,], (0,8], (40,], (33,]
RCT (radiation therapy + adjuvant chemotherapy):
(8,12], (0,5], (30,34], (16,20], (13,], (0,22], (5,8], (13,], (30,36], (18,25]
(24,31], (12,20], (10,17], (17,24], (18,24], (17,27], (11,], (8,21], (17,26], (35,]
(17,23], (33,40], (4,9], (16,60], (33,], (24,30], (31,], (11,], (15,22], (35,39]
(16,24], (13,39], (15,19], (23,], (11,17], (13,], (19,32], (4,8], (22,], (44,48]
(11,13], (34,], (34,], (22,32], (11,20], (14,17], (10,35], (48,]

chemotherapy), where $x_i = (1, x_{1i})$ and regression parameter $\beta = (\beta_0, \beta_1)$, so that the median regression model is $Q_{0\cdot5} = \exp(\beta_0 + \beta_1 x_{1i})$. The patients were observed in irregular intervals, mostly of 4–6 months, for the event of cosmesis. In this study, 46 patients received radiation therapy (RT) and 48 patients received radiation therapy plus adjuvant chemotherapy (RH+CH). The goal of this study is to compare the effect of two treatment arms, RH and RH+CH.

The maximum likelihood estimate of the parameters $(\lambda, \beta_0, \beta_1)$ for the parametric model with Gaussian F_ε and for the semiparametric model with discrete scale-mixture of uniforms for F_ε are given in Table 2. Figure 1 presents the estimated survival functions of two treatment groups under the parametric model. The horizontal lines are nonparametric estimators (Peto 1973) of survival functions, estimated separately for two treatment arms. Based on the maximum likelihood estimator of the semiparametric model, Fig. 2 presents the estimated error density of (3) and the corresponding estimated survival functions of the two groups. This figure also presents the nonparametric estimators for the two treatment arms, for comparison with semi-parametric estimators. We observe that the semiparametric estimators of the survival curves matches the nonparametric estimators better than the match of the parametric estimators in Fig. 1. Using the parametric maximum likelihood estimator and the semiparametric maximum likelihood estimator, Table 3 presents the point and 95 % interval estimates of the medians of the RT ($\exp(\beta_0)$) and RT+CH ($\exp(\beta_0 + \beta_1)$) treatment arms, as well as the ratio of these two medians ($\exp(-\beta_1)$). For both parametric and semiparametric likelihood methods, the interval estimates of the medians and their ratios use the estimated standard errors obtained via the delta-method based on the estimated variance-covariance matrix of the regression parameter estimate $\hat{\beta}$.

For the Bayesian analysis, we assume the parameters to be a priori independent with joint prior

Table 2 Maximum Likelihood estimative of regression parameters β for transformation both-side model

	Estimate	SE	95 % interval estimate
Parametric			
λ	1.629	0.677	(0.302, 2.956)
β_0	3.551	0.133	(3.289, 3.813)
β_1	−0.399	0.183	(−0.759, −0.039)
Semiparametric			
λ	0.968	0.031	(0.908, 1.028)
β_0	3.666	0.072	(3.517, 3.808)
β_1	−0.544	0.101	(−0.743, −0.345)

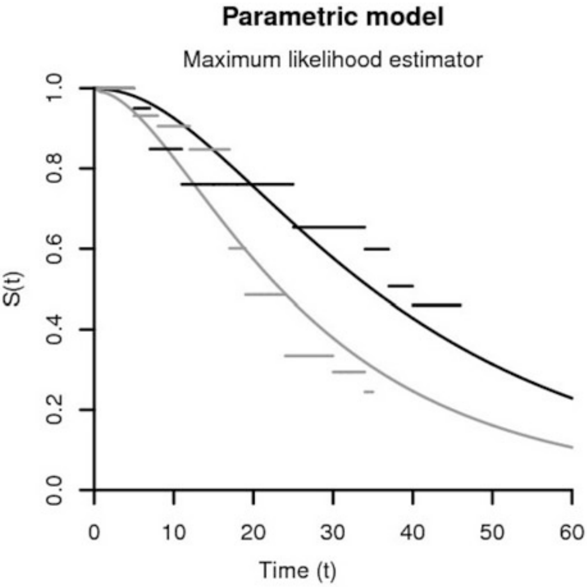


Fig. 1 Parametric maximum likelihood estimator survival functions for two treatment arms, *black* for RT and *gray* for RT+CH, *horizontal lines* are Peto's nonparametric estimators

$$\pi(\lambda, \beta_0, \beta_1, F_\varepsilon) \propto \pi_1(\lambda)\pi_2(\beta_0)\pi_3(\beta_1)\pi_4(F_\varepsilon).$$

It is possible to elicit informative priors for β_0 and β_1 using the prior information, when they are available about the possible support and the prior guess of the median survival times $\exp(\beta_0 + \beta_1)$ and $\exp(\beta_0)$ of two treatment arms. Reviewing the literature, we found no such prior information about these median survival times. We instead used fairly non-informative and skeptical priors, a $N(0, 5^2)$ for β_0, and a $N(0, 1^2)$ for β_1. It is reasonable to assume a priori that the transformation parameter λ has an effective range in the interval $(0, 4)$. Box and Cox (1964) themselves cautioned against using $\lambda >> 3$ due to lack of any reasonable physical interpretation of the model. We suggest using a prior density with mean 1, because $\lambda = 1$ means that no transformation is needed for achieving the symmetry for the log-survival time. In this paper, we are using the gamma prior with mean 1 and variance 1/2.

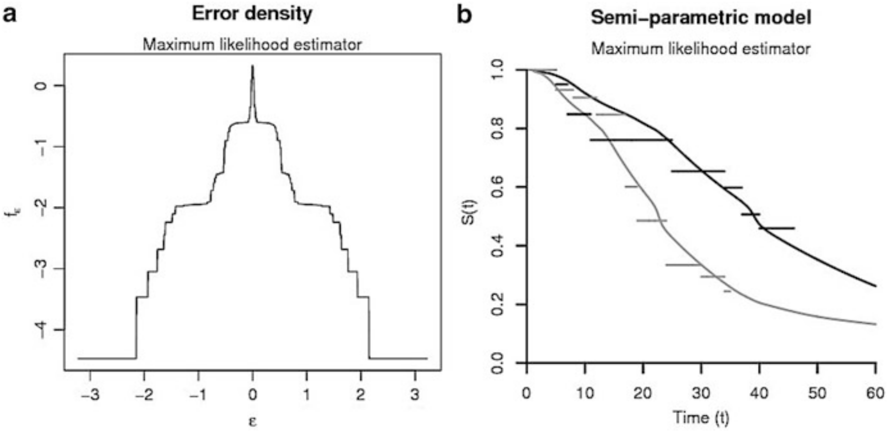


Fig. 2 Semi-parametric maximum likelihood estimator of: (**a**) error density and (**b**) survival functions for two treatment arms, *black* for RT and *gray* for RT+CH, *horizontal lines* are Peto's nonparametric estimators

Table 3 Estimated medians using maximum likelihood estimator

	Point estimate	95 % interval estimate
Parametric		
RT group	34.87	(26.84, 45.32)
RT+CH group	23.39	(18.05, 30.30)
Ratio of medians	1.49	(1.04, 2.13)
Semi-parametric		
RT group	39.10	(33.68, 45.06)
RT+CH group	22.69	(16.12, 31.93)
Ratio of medians	1.72	(1.41, 2.10)

For our analysis, we use the same priors of the parameters (β, λ) for the parametric and semiparametric Bayes analysis.

For the parametric Bayes analysis, we need to specify a prior for the error-variance σ to assign our prior opinion π_4 for Gaussian F_ε. We use a gamma prior with mean 1 and variance 1/2 for σ. For semiparametric Bayes, we use a Gamma distribution with mean 4 and variance 8 as G_0, the prior mean of G. Using Theorem 3, this corresponds to the double-exponential density with scale 0·5 as the prior mean F_0 of F_ε. Also, we use $\alpha = 1$, which means that the precision of the prior opinion about unknown G is very low.

The Bayesian estimates, posterior mean and 95 % credible interval, of the parameters (β_1, λ) for parametric and semiparametric models are given in Table 4. Figure 3 presents the posterior means of the parametric survival functions of two treatment arms. Figure 4 presents the posterior means of the error density F_ε and of the two survival functions under the semiparametric model. Table 5 presents the Bayesian estimates of the median survival times of two groups and the ratio of two medians under these two models. We evaluate the posterior probability of $\beta_1 < 0$

Table 4 Bayesian estimative of transformation both-side model

	Mean	SE	95 % credible interval
Parametric			
λ	1.260	0.300	(0.692, 1.808)
β_0	3.547	0.149	(3.266, 3.821)
β_1	−0.406	0.204	(−0.760, 0.039)
Semi-parametric			
λ	1.054	0.132	(0.807, 1.315)
β_0	3.661	0.158	(3.380, 3.974)
β_1	−0.501	0.202	(−0.920, -0.130)

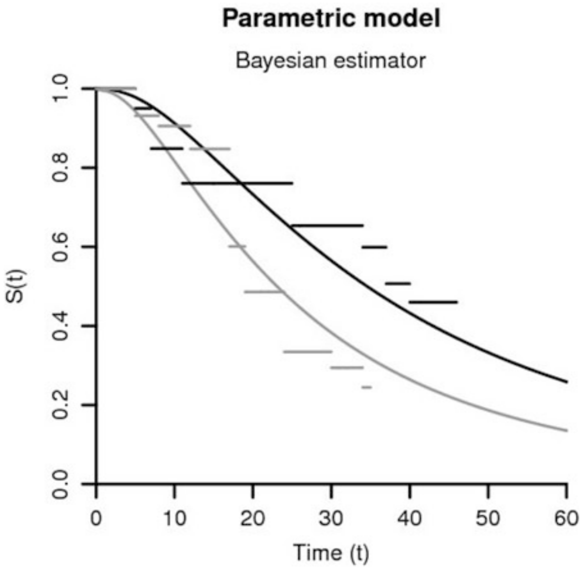


Fig. 3 Parametric posterior mean of survival functions for two treatment arms, *black* for RT and *gray* for RT+CH, *horizontal lines* are Peto's nonparametric estimators

given observed data. For the parametric model, $P(\beta_1 < 0 \mid \mathscr{D}) \cong 0{\cdot}98$, and for the semiparametric model, $P(\beta_1 < 0 \mid \mathscr{D}) \cong 0{\cdot}978$. The convergence diagnostics of the Markov chain Monte Carlo samples were monitored using trace plots and plots of others standard diagnostics.

The results obtained under model (3) can be compared to those obtained under Cox's model $S_1(t \mid x) = \{S_0(t)\}^{\exp(\eta x)}$ only when $S_0(t)$ is exponential. In this special case of Cox's model, the median survival time for covariate x is $Q(x) = \log(2)/\{\nu \exp(\eta x)\}$, where $S_0(t) = \exp(-\nu t)$, and the ratio of medians of two treatment arms, $x = 1$ versus $x = 0$, is equal to $\exp(-\eta)$. The corresponding estimates of $\exp(-\eta)$ obtained by Sinha et al. (1999) and by Finkelstein and Wolfe (1985), are similar, in values, to our maximum likelihood and Bayes estimators of $\exp(\beta_1)$. Overall, there is high posterior evidence to conclude that the median time to observe the cosmetic effects was lower in the patients under the RT+CH treatment than the corresponding median for RT alone.

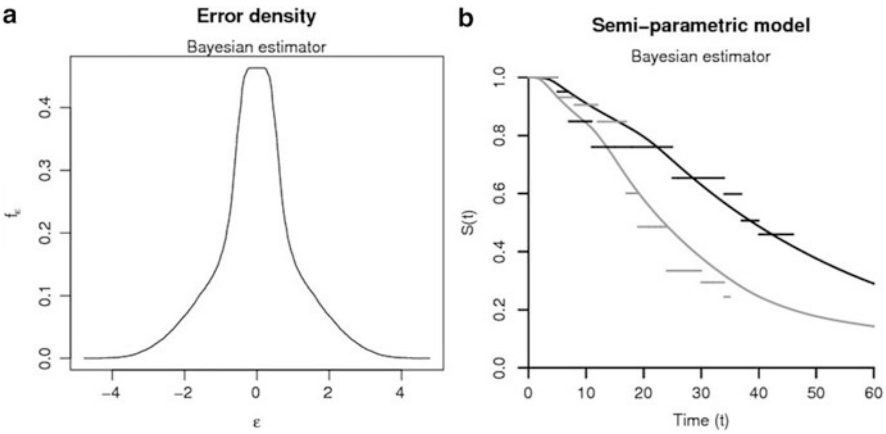


Fig. 4 Semiparametric posterior mean of (**a**) error density and (**b**) survival functions for two treatment arms, *black* for RT and *gray* for RT+CH, *horizontal lines* are Peto's nonparametric estimators

Table 5 Estimated medians using BE

	Point estimate	95 % credible interval
Parametric		
RT group	35.12	(28.20, 45.64)
RT+CH group	23.34	(17.84, 30.16)
Ratio of medians	1.53	(0.96, 2.13)
Semi-parametric		
RT group	39.40	(28.85, 52.40)
RT+CH group	23.73	(18.37, 28.77)
Ratio of medians	1.68	(0.89, 2.22)

For the likelihood analyses, we compare the parametric and semiparametric model estimates of the probability $q_i = P(T \in (A_i, B_i] \mid x_i)$ of the observed data from subject i. In Fig. 5 we present a plot to compare the maximum likelihood estimators for the parametric and semiparametric models. For Bayesian analysis, we also present a similar plot, however, it is based on cross-validated posterior probability $\mathrm{E}[q_i \mid D_{-i}]$ (Gelfand et al. 1992), where D_{-i} is the data based on observed data minus the observation from patient i. For both plots, we see that most of the points, around 70 %, are above the 45° line, implying that the semiparametric model fits the data better under both methods. In both Figs. 2 and 4, the semiparametric estimators of the survival curves under model (3) show good fidelity to the nonparametric estimators. This, supports a better fit of the model (3) for this data compared to the apparent lack of fit of Cox's model, as mentioned in Sinha et al. (1999) among others.

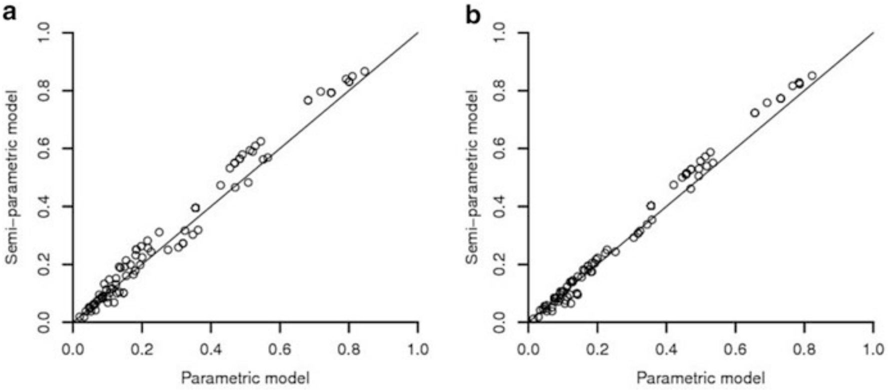


Fig. 5 Comparison between parametric and semi-parametric models: (**a**) maximum likelihood estimator and (**b**) Bayes estimator

5 Final Remarks

The transformation both-side model has some advantages over the other existing methods of inference for interval censored data. The model can focus on the median and quantiles which are more appropriate for continuously monitored studies, than instantaneous risk. The semiparamteric estimation give us a smooth continuous estimated survival functions; The median and any other quantile of the survival time can be obtained from the estimated parameters of the model. In existing quantile regression models (For example, Portnoy 2003), every quantile is assumed to be a linear function $Q_\alpha(x) = \beta_\alpha x$ for every $0 < \alpha < 1$, where $P[T < Q_\alpha(x)] = \alpha$. When x is unbounded, this implies that these linear functions are parallel to each other. In model (3), only one quantile of interest, say the median, is a linear function. The computation of the maximum likelihood estimator involves an iterative algorithm with two simple finite-dimensional maximization steps within each iteration; for semiparametric Bayesian analysis, we only use a Markov chain Monte Carlo technique, implementable via WinBUGS. The hazard functions for this model can be non-monotone. Model (3) can be also written as a location family $Y = \log(T) = \eta + \epsilon$, where $\eta = \beta x$. However, unlike the accelerated lifetime model, the distribution function $F_\varepsilon(\mathrm{g}_\lambda(\epsilon + \eta) - \mathrm{g}_\lambda(\eta))$ of the error ϵ depends on covariate x, a heteroscedastic location family model. The data analysis illustrate the performance of the model, computational and interpretational conveniences as well as the ease of model diagnostics.

For large datasets, the value of K, the number of uniforms used in (8), can be large for the likelihood. For a comparison, in our experience, we found that 7–8 components is large enough to achieve a good approximation for the Sethuraman's construction used in Bayes computation. Thus, using 7–8 components, we have found our approach to be computationally feasible for a large variety of datasets.

References

Ornulf B. Gill R. Keiding N. Andersen, P. (1992). *Statistical Models Based on Countig Processes*. New York: Springer-Verlag.

J. Sun (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. New York: Springer.

D. M. Finkelstein, and R. A. Wolfe (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41,** 933–945.

D. M. Finkelstein, and R. A. Wolfe (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42,** 845–854.

Datta S. Williamson J. Satten (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association* **93,** 318–327.

W. Pan (2000). A multiple imputation approach to cox regression with interval-censored data. *Biometrics* **56,** 199–203.

Johnston G.-Kim H. So, Y. (2010). Analyzing interval-censored data with sas software. *Proceedings of the SAS Global Forum 2010 Conference* **257**.

Chen M.-H. Ghosh S. Sinha (1999). Bayesian analysis and predictive model diagnostics for interval-censored survival data. *Biometrics* **55** 585–590.

Sinha D. Ghosh, S. (2000). Bayesian analysis of interval-censored survival data using penalized likelihood. *Sankhya, Ser. A.* **63** 1–14.

Johnson W. Hanson, T. (2004). A Bayesian semiparametric aft model for interval-censored data. *Journal of Computational and Graphical Statistics* **13** 341–361.

Yang M. Hanson, T. (2007). Bayesian semiparametric proportional odds models. *Biometrics* **63** 88–95.

Piantadosi, S. (2005). *Clinical Trials: A Methodologic Perspective*. Wiley series in probability and statistics. Wiley-Interscience, 2nd ed.

Cheng, S.C., Wei, L.J. and Ying, Z. (1997) Predicting survival probabilities with semiparametric transformation Models. *J. Amer. Statist. Assoc.* **92,** 227–235.

Prentice R. Yang, S. (1999) Semiparametric inference in the proportional odds regression model. *J. Amer. Statist. Assoc.* **94,** 125–136.

Subramanian S. Sun Y. McKeague, I.W. (2001) Median regression and the missing information principle. *J. Nonparametric Statist.* **13,** 709–727.

Tsiatis A. A. Bang, H. (2003) Median regression with censored cost data. *Biometrics* **58,** 643–649.

Portnoy, S. (2003). Censored regression quantiles. *J. Amer. Statist. Assoc.* **98,** 1001–1012.

Huang Y. Peng, L. (2008). Survival analysis with quantile regression models.. *J. Amer. Statist. Assoc.* **103,** 637–649.

Bickel, P.J. and Doksum, K.A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.* **76,** 296–311.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26,** 211–243.

Carroll, R.J. and Ruppert, D. (1984) Power-transformations when fitting theoretical models to data. *J. Amer. Statist. Assoc.* **79,** 321–328.

Fitzmaurice, G.M., Lipsitz, S.R. and Parzen, M. (2007) Approximate median regression via the Box-Cox transformation. *The American Statistician.* **61,** 233–238.

Lin, J., Sinha, D., Lipsitz, S. and Polpo, A. (2012) Semiparametric Bayesian Survival Analysis using Models with Log-linear Median. *Biometrics*. **68,** 1136–1145.

Lin, J., Sinha, D., Lipsitz, S. and Polpo, A. (2013) Bayesian survival analysis using log-linear median regression models. *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association.* 149–158 Springer New York.

Branden K. V. Portnoy S. Neocleous, T. (2006) Correction to censored regression quantiles by s. portnoy, 98. *Journal of the American Statistical Association.* **101**, 860–861.

W. Feller. (1971) An Introduction to Probability Theory and Its Applications. *Wiley.*.

Severini-T. A. Wong, W. H. (1991) On maximum likelihood estimation in infinite dimensional parameter spaces. *The Annals of Statistics*. **19,** 603–632.

Thomas A. Best N. Lunn, D. J. (2000) Winbugs - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*. **10,** 325–337.

A.Y. Khintchine. (1938) On unimodal distributions. *Inst. Mat. Mech. Tomsk. Gos. Univ.*. **2,** 1–7.

Mead R. Nelder, J. A. (1965) A simplex algorithm for function minimization. *Computer Journal*. **7,** 308–313.

van der Vaart A. W. Murphy, S. A. (2001) On profile likelihood. *Journal of the American Statistical Association*. **95,** 449–465.

Sethuraman, J. (1994). Constructive definition of dirichlet priors. *Statistica Sinica* **4,** 639–650.

Dey D. Chang H. Gelfand, A. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *In Bayesian Statistics 4, J. Bernardo, J. Berger, A. Dawid and A. Smith, eds. Oxford University Press.*

T.S. Ferguson (1973). Bayesian analysis of some nonparametric problems. The Annals of Statistics, **1,** 209–230.

R. Peto (1973). An experimental survival curve for interval-censored data. Journal of the Royal Statistical Society, Ser. C (Applied Statistics), **22,** 86–91.

Explained Variation for Correlated Survival Data Under the Proportional Hazards Mixed-Effects Model

Gordon Honerkamp-Smith and Ronghui Xu

Abstract Measures of explained variation are useful in scientific research, as they quantify the amount of variation in an outcome variable of interest that is explained by one or more other variables. We develop such measures for correlated survival data, under the proportional hazards mixed-effects model (PHMM). Since different approaches have been studied in the literature outside the classical linear regression model, we investigate four sample-based measures that estimate three different population coefficients. We show that although the three population measures are not the same, they reflect similar amounts of variation explained by the predictors. Among the four sample-based measures, we show that the first one (R^2) which is the simplest to compute, is also consistent for the first population measure (Ω^2) under the usual asymptotic scenario when the number of clusters tends to infinity; the other three sample-based measures, on the other hand, all require that in addition the cluster sizes be large. We study the properties of the measures through simulation studies. We illustrate their usage on a multi-center clinical trial data set.

Keywords Clustered survival data • Explained randomness • Multi-center clinical trial • Recurrent events

1 Introduction

Correlated survival data arise in many areas of biomedical applications. They arise in multi-center clinical trials where, despite rigorously designed protocols, complex procedures and different clinical practices may lead to different treatment effects at different centers. Recurrent events are another type of correlated survival data,

G. Honerkamp-Smith
Department of Mathematics, University of California, San Diego, CA, USA

R. Xu (✉)
Department of Mathematics, University of California, San Diego, CA, USA

Department of Family Medicine and Public Health, University of California,
San Diego, CA, USA

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_14

though with their specific chronological orders. Genetic studies, often by design, recruit groups of subjects who are family members and share the same genetic or environmental factors. Like for independently and identically distributed (i.i.d.) data, often we would like to be able to quantify the amount of variation in the correlated outcomes that is explained by the predictors, which is an important attribute of any regression model.

The R^2 coefficient of determination in classical linear regression is the definitive solution to such a need. For correlated outcomes data, random effects models (sometimes called variance components models) are a natural way of decomposing the variation in the outcomes into different components (Xu 2003). As an example of application in genetic epidemiology, it is common to decompose the variation in a disease outcome into contributions from genetic, environmental, and residual components (Sneider et al. 1999; Liu et al. 2004a, 2005), all expressed as percentages that add up to one.

To this time, much attention has been given to developing measures of explained variation in the presence of right-censored survival data. Many of the early proposals were based on extensions of different yet equivalent definitions of the R^2 coefficient of determination under the multiple linear regression model (Kvalseth 1985). These extensions are not the same outside of the normal linear model. A comparison of the early proposals can be found in Schemper and Stare (1996). Proposals have also been made in the literature based on computationally intensive methods such as multiple imputation of the censored observations (Schemper and Kaider 1997). More recently Heller (2012) proposed a measure of explained risk (instead of variation) under the Cox model, and Preseley et al. (2011) applied some of these measures to surrogate evaluation. A recent discussion of the related concepts and recommendations can be found in O'Quigley and Xu (2012).

For analyzing correlated survival data, mixed-effects models have been proposed that specify the correlation structure within the outcomes, as well as to correlate with the predictors. In this paper we consider the proportional hazards mixed-effects model (PHMM) (Ripatti and Palmgren 2000; Vaida and Xu 2000). This model encompasses the commonly known frailty model, which contains random intercepts but not random effects on arbitrary covariates. Under the PHMM we aim to define both population measures of explained variation, as well as their sample based estimates. We explore a couple of commonly used approaches, which include a direct decomposition of the variance, a ratio of sums of squares, and an information theoretical measure that is easily computed by transforming the likelihood ratio statistic. In the following we will first recall details of the PHMM and related quantities that will be used to define the measures of explained variation.

1.1 Model and Notation

The PHMM extends the Cox proportional hazards model by including a vector of random effects terms in the log relative risk:

$$\lambda_{ij}(t) = \lambda_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{Z}_{ij} + \boldsymbol{b}_i'\boldsymbol{W}_{ij}), \quad i = 1, \ldots, m; j = 1, \ldots, n_i. \tag{1}$$

Here, $\lambda_{ij}(t)$ is the hazard function of the j-th observation in the i-th cluster of size n_i, $\boldsymbol{\beta}$ is the vector of fixed effects, $\boldsymbol{b}_i$ is a vector of random effects associated with cluster i, and $\boldsymbol{Z}_{ij}$ and $\boldsymbol{W}_{ij}$ are covariate vectors corresponding to the fixed and random effects, respectively. The event time T_{ij} may be right-censored; we observe $X_{ij} = \min(T_{ij}, C_{ij})$, where C_{ij} is the potential censoring time. Let $\delta_{ij} = \mathbf{1}\{T_{ij} \leq C_{ij}\}$, and $Y_{ij}(t) = \mathbf{1}\{X_{ij} \geq t\}$ be the "at-risk" indicator at time t. It is usually assumed that for every covariate with a random effect there is also a corresponding fixed effect, so that $\boldsymbol{W}_{ij}$ is a subset of $\boldsymbol{Z}_{ij}$ except possibly for a '1' in the first entry that models the random cluster effect on the baseline hazard. In general the random effects can be seen as cluster by covariate interactions (Vaida and Xu 2000). Thus, the data consist of the triples $(X_{ij}, \delta_{ij}, \boldsymbol{Z}_{ij})$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. The random effects $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m$ are independent of each other, and assumed to be $N(0, \Sigma)$; they are also assumed to be independent of the covariates $\boldsymbol{Z}$.

The following quantities under the PHMM are relevant to our development later. Conditional on the $\boldsymbol{b}_i$'s, at each time t we have a probability distribution on the set of subjects at risk, given by:

$$\pi_{ij}(t; \boldsymbol{\beta}, \boldsymbol{b}) = \frac{Y_{ij}(t)\exp(\boldsymbol{\beta}'\boldsymbol{Z}_{ij} + \boldsymbol{b}_i'\boldsymbol{W}_{ij})}{\sum_{k,l} Y_{kl}(t)\exp(\boldsymbol{\beta}'\boldsymbol{Z}_{kl} + \boldsymbol{b}_k'\boldsymbol{W}_{kl})}. \tag{2}$$

The term $\pi_{ij}(t; \boldsymbol{\beta}, \boldsymbol{b})$ can be interpreted as the probability that the j-th subject in cluster i fails at time t given the risk set and that exactly one failure occurs at that time. Evaluating π_{ij} at time $t = X_{ij}$ and taking the product of such terms over the observed failure times ($\delta_{ij} = 1$) forms the partial likelihood conditional on the $\boldsymbol{b}_i$'s:

$$L_p(\boldsymbol{\beta}, \boldsymbol{b}) = \prod_{ij:\delta_{ij}=1} \pi_{ij}(X_{ij}; \boldsymbol{\beta}, \boldsymbol{b}). \tag{3}$$

The above was used in Ripatti and Palmgren (2000) to form the penalized partial likelihood under the PHMM. It is shown that the discrete probability distribution $\{\pi_{ij}(t; \boldsymbol{\beta}, \boldsymbol{b})\}_{i=1\ldots m, j=1\ldots n_i}$ converges weakly to the conditional distribution of $\boldsymbol{Z}$ given $T = t$ and the $\boldsymbol{b}_i$'s, in the same way that an empirical distribution converges to the underlying distribution function (Xu and Gamst 2007). Under the classic Cox model this conditional distribution has been used to construct time-dependent ROC curves (Heagerty and Zheng 2005).

The model parameters $\theta = (\boldsymbol{\beta}, \Sigma, \lambda_0)$ can be consistently estimated by the nonparametric maximum likelihood estimator (NPMLE), which has been shown to have optimal asymptotic and numerical properties (Gamst et al. 2009). The NPMLE can be computed using an MCEM algorithm, and is available in the R package 'phmm'. At the convergence of the algorithm, the posterior distribution $p(\boldsymbol{b}_i | \boldsymbol{y}_i, \hat{\boldsymbol{\theta}})$ of $\boldsymbol{b}_i$, where $\boldsymbol{y}_i$ represents the observed data from cluster i, can be used to produce empirical Bayes "estimates" of the random effects. In doing so we are viewing the realized values of the $\boldsymbol{b}_i$'s like parameters, estimated via a degree of shrinkage; this notion is closely related to the conditional inference discussed in Vaida and Blanchard (2005) and Donohue et al. (2011). We will make use of the empirical Bayes estimates $\hat{\boldsymbol{b}}_i = \mathrm{E}\left(\boldsymbol{b}_i|\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}\right)$ when defining some of the measures below.

Finally, model (1) is known to be equivalent to the linear transformation mixed-effects model (Xu and Gamst 2007)

$$g(T_{ij}) = -(\boldsymbol{\beta}'\boldsymbol{Z}_{ij} + \boldsymbol{b}_i'\boldsymbol{W}_{ij}) + \epsilon_{ij}, \tag{4}$$

where $g(\cdot) = \log \Lambda_0(\cdot)$ is a monotone transformation, and ϵ has the fixed (and known) extreme value distribution with variance $\pi^2/6$. The general semiparametric transformation model with mixed effects in the form of (4) was considered in Zeng and Lin (2007).

In the next section, we present measures of explained variation, both population and sample based, and discuss some of their properties. Simulation studies are carried out in Sect. 3, and the measures are applied to real data in Sect. 4.

2 Measures of Explained Variation

In the context of the semiparametric regression models like (1) or (4), the specified part of the model only concerns the prediction of the ranks of the T's given the $\boldsymbol{Z}$'s. The actual scale of the failure times as reflected in the observed data is not modeled, and is estimated by the nonparametric baseline hazard or the nonparametric transformation. In addition, in the presence of clustering in the data, the analysis is often concerned with how much variation is explained by the covariates or even the clustering itself.

2.1 *Explained Variation Ω^2 and Its Estimates*

The explained variation in a response A by its predictors $\boldsymbol{Z}$ can be defined based on the well-known formula $\mathrm{Var}(A) = E\{\mathrm{Var}(A|\boldsymbol{Z})\} + \mathrm{Var}\{E(A|\boldsymbol{Z})\}$ (O'Quigley and Xu 2012). The first term in the decomposition can be seen as the expected residual variance in A after using $\boldsymbol{Z}$ to 'explain' A, and the second term as the variability explained by the conditional distribution of A given $\boldsymbol{Z}$, often modeled by the regression. Under model (4) we consider $A = g(T)$. The proportion of explained variation is then

$$\Omega^2 = 1 - \frac{E\{\mathrm{Var}\,(g(T))\,|\boldsymbol{Z}\}}{\mathrm{Var}\,(g(T))} = 1 - \frac{\pi^2/6}{\mathrm{Var}\left(\boldsymbol{\beta}'\boldsymbol{Z} + \boldsymbol{b}'\boldsymbol{W}\right) + \pi^2/6}, \tag{5}$$

where $\pi^2/6$ is the error variance, and $\boldsymbol{Z}$, $\boldsymbol{b}$, $\boldsymbol{W}$ are generic versions of $\boldsymbol{Z}_{ij}$, $\boldsymbol{b}_i$, $\boldsymbol{W}_{ij}$, i.e. random variables (or vectors) with the same distributions. Note that (5) was used in the genetic epidemiology literature to quantify the genetic versus environmental contributions disease onset (Liu et al. 2004a,b, 2005).

Evidently, estimation of Ω^2 can be accomplished by estimating the variance of the linear predictor $\eta = \boldsymbol{\beta}'\boldsymbol{Z} + \boldsymbol{b}'\boldsymbol{W}$. We will consider two ways of estimating this quantity. We first express η as the inner product of two independent random vectors; the variance can then be given in terms of their moments, and hence a final estimate is obtained by estimating these moments. For the purpose of notation, we will assume that the model includes a random intercept; the computation is slight simpler without one. We assume that the vector $\boldsymbol{W}$ consists of a 1 followed by the first p coordinates of $\boldsymbol{Z} \in \mathbb{R}^{p+q}$. Define the following vectors in $\mathbb{R}^{p+q+1}$:

$$\tilde{\boldsymbol{\beta}} = \begin{pmatrix} 0 \\ \boldsymbol{\beta} \end{pmatrix}, \; \widetilde{\boldsymbol{Z}} = \begin{pmatrix} 1 \\ \boldsymbol{Z} \end{pmatrix}, \text{ and } \tilde{\boldsymbol{b}} = \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{0} \in \mathbb{R}^q$ is a vector of zeros. With $\boldsymbol{U} = \tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{b}}$ and $\boldsymbol{V} = \widetilde{\boldsymbol{Z}}$ we have $\boldsymbol{\eta} = \boldsymbol{U}'\boldsymbol{V}$. The expectations of $\boldsymbol{U}$ and $\boldsymbol{V}$ are

$$\mathrm{E}\,(\boldsymbol{U}) = \tilde{\boldsymbol{\beta}} = \begin{pmatrix} 0 \\ \boldsymbol{\beta} \end{pmatrix}, \; \mathrm{E}\,(\boldsymbol{V}) = \begin{pmatrix} 1 \\ \boldsymbol{\mu}_Z \end{pmatrix},$$

where $\boldsymbol{\mu}_Z$ denotes the expectation of $\boldsymbol{Z}$. The covariance matrices are

$$\mathrm{Var}\,(\boldsymbol{U}) = \mathrm{Var}\left(\tilde{\boldsymbol{b}}\right) = \begin{pmatrix} \Sigma_b & \\ & O_{q \times q} \end{pmatrix}, \text{ and } \mathrm{Var}\,(\boldsymbol{V}) = \mathrm{Var}\left(\widetilde{\boldsymbol{Z}}\right) = \begin{pmatrix} 0 & \\ & \Sigma_Z \end{pmatrix}$$

where $O_{a \times b}$ is an $a \times b$ matrix of zeroes. Brown and Rutemiller (1977) provides a formula for the variance of the inner product of two independent random vectors: $\mathrm{Var}\left(\boldsymbol{U}'\boldsymbol{V}\right) = \mu_U'\Sigma_V\mu_U + \mu_V'\Sigma_U\mu_V + \mathrm{tr}(\Sigma_U\Sigma_V)$, where μ and Σ with a subscript denote the expectation and covariance matrix of the random vector indicated. Define $\boldsymbol{b}_1$ by $\boldsymbol{b} = (b_0, \boldsymbol{b}_1')'$, let $\boldsymbol{Z}_1$ be the first p components of $\boldsymbol{Z}$, and let $\tilde{\mu}_{Z_1} = (1, \mu_{Z_1}')'$. Then the variance of η is $\mathrm{Var}\,(\eta) = \boldsymbol{\beta}'\Sigma_Z\boldsymbol{\beta} + \tilde{\mu}_{Z_1}'\Sigma_b\tilde{\mu}_{Z_1} + \mathrm{tr}(\Sigma_{b_1}\Sigma_{Z_1})$. Therefore Ω^2 is a function of the population parameters $\boldsymbol{\beta}$, $\boldsymbol{\mu}_Z$, Σ_Z, and Σ_b:

$$\Omega^2 = 1 - \frac{\pi^2/6}{\boldsymbol{\beta}'\Sigma_Z\boldsymbol{\beta} + \tilde{\mu}_{Z_1}'\Sigma_b\tilde{\mu}_{Z_1} + \mathrm{tr}(\Sigma_{b_1}\Sigma_{Z_1}) + \pi^2/6}. \tag{6}$$

Parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\Sigma}_b$ can be computed under the PHMM using the previously mentioned R package ‘*phmm*’. Using the sample estimators for the mean and variance of $\boldsymbol{Z}$, we can estimate Ω^2 by

$$R^2 = 1 - \frac{\pi^2/6}{\hat{\boldsymbol{\beta}}'\hat{\Sigma}_Z\hat{\boldsymbol{\beta}} + \hat{\tilde{\mu}}'_{Z_1}\hat{\Sigma}_b\hat{\tilde{\mu}}_{Z_1} + \mathrm{tr}(\hat{\Sigma}_{b_1}\hat{\Sigma}_{Z_1}) + \pi^2/6}. \tag{7}$$

Alternatively, the realized values of the $\boldsymbol{b}_i$'s can be 'estimated' using empirical Bayes method. Combined with the estimate $\hat{\boldsymbol{\beta}}$ of the fixed effects, the linear predictor for each observation can be estimated as $\hat{\eta}_{ij} = \hat{\boldsymbol{\beta}}'\boldsymbol{Z}_{ij} + \hat{\boldsymbol{b}}'_{ij}\boldsymbol{W}_{ij}$. A second estimate of Ω^2 is then based on the sample variance of the $\hat{\eta}_{ij}$'s:

$$R_1^2 = 1 - \frac{\pi^2/6}{\sum_{i,j}(\hat{\eta}_{ij} - \bar{\hat{\eta}})^2/(N-1) + \pi^2/6}, \tag{8}$$

where $N = \sum_{i=1}^m n_i$ is the total number of observations. It is known that for the $\boldsymbol{b}_i$'s to be well estimated, the cluster sizes n_i need to be reasonably large, and this is when we expect (8) to be a reasonable estimate of Ω^2.

2.2 A Sum of Squares Approach

In classical linear regression the R^2 can be expressed as a ratio of sums of squared residuals. A well-known type of residual under the proportional hazards regression is the Schoenfeld residual (Schoenfeld 1982), which has been used in O'Quigley and Flandre (1994) and O'Quigley and Xu (2012) to define R^2 measures. O'Quigley and Xu (2012) also extended the Schoenfeld residuals to 'residuals' of the prognostic index. Under the univariate Cox model, η has a one-to-one correspondence to the covariate Z, assuming that $\beta \neq 0$. We now extend this method to the PHMM setting. Note that under model (1) or, equivalently model (4), the predicted ranks of T have a one-to-one correspondence to the prognostic index $\eta = \boldsymbol{\beta}'\boldsymbol{Z} + \boldsymbol{b}'\boldsymbol{W}$. In this sense η is like a 'surrogate' for the actual, possibly censored outcome T. This fact has been used in the prediction context by, for example Huang and Harrington (2002), to select the penalty parameters.

In order to define the relevant residuals, we first need to define the expected prognostic index at a given failure time t, using the probability distribution defined in (2):

$$\mathrm{E}^{\pi}_{\boldsymbol{\beta},\boldsymbol{b}}(\eta; t) = \sum_{i,j} \eta_{ij}\pi_{ij}(t; \boldsymbol{\beta}, \boldsymbol{b}) = \sum_{i,j} \frac{Y_{ij}(t)\eta_{ij}\exp\{\eta_{ij}\}}{\sum_{k,l} Y_{kl}(t)\exp\{\eta_{kl}\}}. \tag{9}$$

Here again we view the realized values of $\boldsymbol{b}$ as parameters, to be estimated via the empirical Bayes shrinkage under the PHMM. At each failure time, we can then compare the value of η predicted by the model as in (9) with the one actually observed. Having estimated $\boldsymbol{\beta}$ and the $\boldsymbol{b}_i$'s, the estimated prognostic index for the ij^{th} observation is $\hat{\eta}_{ij} = \hat{\boldsymbol{\beta}}'\boldsymbol{Z}_{ij} + \hat{\boldsymbol{b}}'_i\boldsymbol{W}_{ij}$. This gives the residuals

$$r_{ij}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{b}}) = \hat{\eta}_{ij} - \mathrm{E}^{\pi}_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{b}}}(\hat{\eta};X_{ij}), \tag{10}$$

whenever $\delta_{ij} = 1$.

To form what is equivalent to a total sum of squares, we consider a ‘null’ model in order to contrast with the full model in question. When the interest lies in quantifying the amount of variation in the survival that is explained by both the covariates and the clustering itself, the latter modeled by b_0, the corresponding null model is given by $\boldsymbol{\beta} = 0$ and $\boldsymbol{b} = 0$, and the hazard function is simply $\lambda(t|\boldsymbol{Z},\boldsymbol{b}) = \lambda_0(t)$. Let $\mathcal{R}(t)$ be the risk set at time t and $|\mathcal{R}(t)| = \sum_{i,j} Y_{ij}(t)$ its size. Under this null model all subjects in the risk set have the same probability for failure:

$$\pi_{ij}(t) = \frac{Y_{ij}(t)}{\sum_{k,l} Y_{kl}(t)} = \frac{Y_{ij}(t)}{|\mathcal{R}(t)|},$$

and the expected η at time t is the just the simple average over the risk set:

$$\overline{\eta}(t) = \mathrm{E}^{\pi}_{0}(\eta;t) = \frac{\sum_{i,j} Y_{ij}(t)\eta_{ij}(t)}{|\mathcal{R}(t)|}. \tag{11}$$

The ‘null’ residuals are then

$$r^0_{ij} = r_{ij}(0,0) = \hat{\eta}_{ij}(X_{ij}) - \overline{\hat{\eta}}(X_{ij}). \tag{12}$$

We can now define the coefficient of explained variation using the residual sum of squares under the full and the null models:

$$R^2_2 = 1 - \frac{\sum_{i,j} \delta_{ij} r_{ij}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{b}})^2}{\sum_{i,j} \delta_{ij}(r^0_{ij})^2}.$$

2.3 *Explained Randomness*

Kent (1983) developed a general notion of explained randomness based on the Kullback-Leibler (KL) information gain, and it has been applied to the proportional hazards regression model (Kent and O’Quigley 1988; Xu and O’Quigley 1999; O’Quigley et al. 2005). A commonly encountered pitfall in the literature when defining such a measure, sometimes called the generalized R^2 (Cox and Snell 1989), is ignoring the original definition based on the KL information and simply taking an *ad hoc* transformation of the likelihood ratio statistics; this in particular can lead to erroneous definitions in the presence of censored data (O’Quigley et al. 2005). Here we develop the explained randomness measure for the PHMM.

As discussed in Kent (1983) when using the explained randomness to capture the dependence between two random variables, there is a certain degree of symmetry in using the conditional distribution of one variable given the other, or vice versa. In the special case of bivariate normal, no matter which way one conditions, the explained randomness is equal to the correlation coefficient squared. In the context of the semiparametric proportional hazards regression, predicting ranks of the T's given the $\boldsymbol{Z}$'s is equivalent to predicting the $\boldsymbol{Z}$'s given the T's (O'Quigley and Xu 2012). In this way, it is natural to consider the conditional distribution of $\boldsymbol{Z}$ given T; this is also consistent with the partial likelihood inference procedure, as well as the residuals considered in the Sect. 2.2.

As before θ denotes the unknown parameters under the PHMM. The KL information is

$$I(\theta) = \mathrm{E}\left(\log\{f(\boldsymbol{Z}|T, \boldsymbol{b}; \theta)\}\right), \tag{13}$$

where $f(\cdot)$ is the conditional density or probability function of $\boldsymbol{Z}$ given T and $\boldsymbol{b}$, and the expectation is taken with respect to the true underlying distribution. For two nested models indexed by $\theta \in \Theta_0 \subset \Theta_1$, let $\theta_i = \text{argmax}\{I(\theta); \theta \in \Theta_i\}$ $(i = 0, 1)$, and $\Gamma = 2\{I(\theta_1) - I(\theta_0)\}$. If Θ_0 is the subset of model distributions for which T and $\boldsymbol{Z}$ are independent, we can think of Γ as measuring the information gained from modeling dependence. In that case Kent (1983) called $\exp\{-2I(\theta_0)\}$ the total randomness in $\boldsymbol{Z}$, and $\exp\{-2I(\theta_1)\}$ the residual randomness of $\boldsymbol{Z}$ given T. The proportion of explained randomness is then

$$\rho^2 = 1 - \frac{\exp\{-2I(\theta_1)\}}{\exp\{-2I(\theta_0)\}} = 1 - \exp(-\Gamma). \tag{14}$$

The expectation in (13) is typically unknown, but can be estimated by the empirical distribution of the data in general (Kent 1983). For example Γ can be estimated by $1/n$ times the likelihood ratio statistics for testing Θ_1 versus Θ_0 for a random sample of size n. As described in the Introduction, under the PHMM the conditional distribution of $\boldsymbol{Z}$ given T and $\boldsymbol{b}$ is estimated by $\{\pi_{ij}(t; \boldsymbol{\beta}, \boldsymbol{b})\}_{i=1\ldots m, j=1\ldots n_i}$. From (3) the log partial likelihood conditional on $\boldsymbol{b}$ is

$$\log L_p(\boldsymbol{\beta}, \boldsymbol{b}) = \sum_{\delta_{ij}=1} \left\{ \boldsymbol{\beta}'\boldsymbol{Z}_{ij} + \boldsymbol{b}_i'\boldsymbol{W}_{ij} - \log\left(\sum_{(kl)\in\mathcal{R}(X_{ij})} e^{\boldsymbol{\beta}'\boldsymbol{Z}_{ij} + \boldsymbol{b}_i'\boldsymbol{W}_{ij}} \right) \right\}. \tag{15}$$

Under the null model $\boldsymbol{\beta} = 0, \boldsymbol{b} = 0$, and $\lambda(t|\boldsymbol{Z}, \boldsymbol{b}) = \lambda_0(t)$ as in Sect. 2.2. The log partial likelihood becomes

$$\log L_p^0 = -\sum_{\delta_{ij}=1} \log|\mathcal{R}(X_{ij})|. \tag{16}$$

In the presence of right-censoring, the effective sample size is $K = \sum_{i,j} \delta_{ij}$ which is the total number of events; K is also the number of terms in the log partial likelihood. Using an empirical distribution assigning mass $1/K$ to each observed failure to further approximate the expectation in (13) (O'Quigley et al. 2005), the estimated information gain is then

$$\widehat{\Gamma} = \frac{2}{K}\left\{L_p(\boldsymbol{\beta}, \boldsymbol{b}) - L_p^0\right\}. \tag{17}$$

Our measure of explained randomness is based on this information gain:

$$\hat{\rho}^2 = 1 - \exp(-\hat{\Gamma}). \tag{18}$$

3 Simulation

In Fig. 1 we compare the four proposed measures and their population values. For a fixed value of β, the survival times T_{ij} were generated according to $\lambda_{ij}(t) = \exp(\beta Z_{ij} + b_{0i} + b_{1i} Z_{ij})$, where $Z_{ij} \sim N(0.5, 0.25)$, $b_{0i}, b_{1i} \sim N(0, 0.25)$. Independent censoring times C_{ij} were generated from a uniform distribution on the interval $(0, \tau)$, where τ was chosen so that there was about 25 % censoring. The PHMM was then fit to the dataset using the *phmm*() function in the R package 'phmm'.

While the population value for R^2 and R_1^2 is Ω^2 given in (5), the population value for R_2^2 as well as ρ^2 do not have closed-form expressions, and are obtained by using Monte Carlo simulation with a large sample size of 200 clusters with 50 observations in each cluster (200×50). These three population measures are marked by points with a square, circle, or triangle. The fact that they increase with $|\beta|$ translates to improved predictive capability as $|\beta|$ increases. Note that even

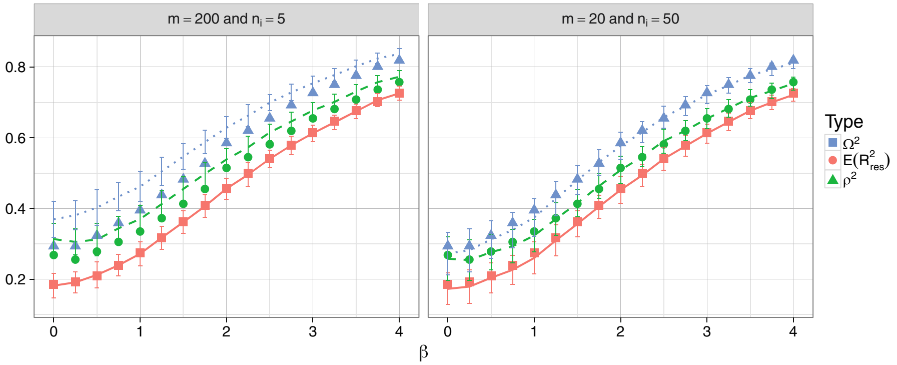


Fig. 1 Ω^2, R^2 and R_1^2 increase with $|\beta|$; $Z \sim$ N(0.5, 0.25), Var(b_0) = Var(b_1) = 0.25

when $\beta = 0$, the measures of explained variation are non-zero. This is because the model still retains the random effects, which explain part of the variation in the data: $\lambda_{ij}(t) = \lambda_0(t)\exp(b_{0i} + b_{1i}Z_{ij})$.

From the figure it is clear that the three population measures are different quantities; they do, however, reflect similar strengths of predictability in our opinion, differing from each other by at most 10 % in all cases. The sample-based measures are plotted using different line types, as noted in the figure legend. In comparing the left (200 × 5) versus the right (20 × 50) panels of Fig. 1, we see how the sample sizes affect the accuracy of these measures in estimating their population equivalents. In particular, we see that in the left panel R^2 accurately estimates Ω^2, while the other three measures, all relying on the estimated $\hat{\boldsymbol{b}}_i$'s, are not good estimates of their population equivalents due to the small cluster size of 5. On the other hand, in the right panel R_1^2, R_2^2 and $\hat{\rho}^2$ are much closer to their population equivalents, while R^2 is a bit less accurate in estimating Ω^2 than in the left panel due to the smaller number of clusters 20. Note that the number of clusters is the sample size that affects the frequentist model parameters, while the cluster size affects the accuracy of $\hat{\boldsymbol{b}}_i$'s.

4 An Application

In Vaida and Xu (2000) illustrated the application of the PHMM using a multi-center clinical trial in lung cancer conducted by the Eastern Cooperative Oncology Group (E1582). There were 579 patients from 31 institutions randomized to one of two chemotherapy regimens. The overall survival time was observed along with five relevant binary baseline covariates: treatment, presence of bone metastasis, presence of liver metastasis, ambulatory performance status, and weight loss prior to treatment. Gray (1995) developed tests for variation across groups in survival data and showed that, for this dataset, there is significant variation by institution in the treatment effect. Vaida and Xu (2000) and Xu et al. (2009) fitted the PHMM to the data, and discovered random effects of bone metastases, which had even larger variance than the random effects for treatment. In the Bayesian variable selection context (Dunson and Chen 2004) concluded that after accounting for the random bone metastases effects, there was no direct evidence of institutional variation in treatment effects. This was then followed by a correspondence from Gray (2006) and a further discussion in Lee et al. (2014).

In the following we consider the explained variation for this data set. We first consider univariate analyses allowing for random effects if necessary. In Table 1 we present the PHMM fits for treatment and bone metastases separately, each with a random effect. The initial fits of the PHMM to the other three covariates separately all had their variances of the random effects converging to zero during the EM iterations (Vaida and Xu 2000), and is hence presented with results from the regular Cox model fits without random effects. From the table we see that with or without the random effects, each covariate only explains a small percentage of variation in overall survival, indicating the each binary variable alone does not make

Table 1 Univariate fits of the lung cancer data; σ^2 is the variance of the random slope

Covariate	β	σ^2	R^2	R_1^2	R_2^2	$\hat{\rho}^2$
Trt	−0.28 (0.10)	0.05 (0.03)	0.03	0.02	0.07	0.06
Bone	0.35 (0.14)	0.19 (0.12)	0.05	0.04	0.07	0.09
Liver	0.45 (0.09)	–	0.03	0.03	0.05	0.05
PS	−0.58 (0.10)	–	0.03	0.03	0.06	0.05
Wtlss	0.27 (0.09)	–	0.01	0.01	0.02	0.02

Table 2 Multivariate fits of the lung cancer data; all five fixed covariate effects are included

Random effects	σ^2	R^2	R_1^2	R_2^2	$\hat{\rho}^2$
None	–	0.09	0.09	0.13	0.13
Treatment	0.07 (0.05)	0.11	0.11	0.16	0.17
Bone	0.14 (0.08)	0.13	0.11	0.17	0.18
Treatment	0.05 (0.08)	0.13	0.12	0.19	0.21
+ Bone	0.13 (0.12)				

a good predictor for survival, which is probably the case in reality. In comparing the four measures, we see that R_2^2 and $\hat{\rho}^2$ gave slightly higher values than R^2 and R_1^2, consistent with our numerical findings of the previous section.

In Table 2 we incorporate all five covariates, and allow none, treatment only, bone metastases only, or both treatment and bone metastases random effects. For ease of discussion we mainly focus on the R^2 values, although in our opinion all four measures reflect a similar degree of predictability by the covariates, with the last two having slightly higher values. Note the data structure is such that each institution varies between a size of 1–50 patients, with an average of just under 20 patients per institution. It is seen that in terms of explained variation, the five covariates together explain about 10 % of the variation in overall survival, with allowing for random effects explaining a couple of percentage points. It is also seen that the random bone metastases effect explains just a little bit more variation than the random treatment effect, and that adding the random treatment effect to the random bone metastases effect does not appear to explain much additional variation. In Lee et al. (2014) the authors also discussed the distinction between a relatively weak random effect (treatment) and a relatively strong random effect (bone metastases), and their impact on Bayesian variable selection. Our observation here appears consistent with those discussions.

References

Xu R. Measuring explained variation in linear mixed effects models. *Statistics in Medicine* 2003; **22**:3527–3541.

Sneider H, Boomsma DI, van Doornen LJP, Neale MC. Bivariate genetic analysis of fasting insulin and glucose levels. *Genetic Epidemiology* 1999; **16**:426–446.

Liu I, Blacker DL, Xu R, Fitzmaurice G, Lyons MJ, Tsuang MT. Genetic and environmental contributions to the development of alcohol dependence in male twins. *Archives of General Psychiatry* 2004a; **61**:897–903.

Liu I, Xu R, Blacker DL, Fitzmaurice G, Lyons MJ, Tsuang MT. The application of a random effects model to censored twin data. *Behavior Genetics* 2005; **35**:781–789.

Kvalseth TO. Cautionary note about R^2. *The American Statistician* 1985; **39**:279–285.

Schemper M, Stare J. Explained variation in survival analysis. *Statistics in Medicine* 1996; **15**:1999–2012.

Schemper M, Kaider A. A new approach to estimate correlation coefficients in the presence of censoring and proportional hazards. *Computational Statistics and Data Analysis* 1997; **23**: 467–476.

Heller G. A measure of explained risk in the proportional hazards model. *Biostatistics* 2012; **13**:315–325.

Preseley A, Tilahun A, Alonso A, Molenberghs G. An information-theoretic approach to surrogate-marker evaluation with failure time endpoints. *Lifetime Data Analysis* 2011; **17**:195–214.

O'Quigley J, Xu R. Explained variation and explained randomness for proportional hazards models. *Handbook of Statistics in Clinical Oncology (3rd Ed.), ed. Crowley and Hoering* 2012; :487–503.

Ripatti S, Palmgren J. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* 2000; **56**:1016–1022.

Vaida F, Xu R. Proportional hazards model with random effects. *Statistics in Medicine* 2000; **19**:3309–3324.

Xu R, Gamst A. On proportional hazards assumption under the random effects models. *Lifetime Data Analysis* 2007; **13**:317–332.

Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005; **61**:92–105.

Gamst A, Donohue M, Xu R. Asymptotic properties and empirical evaluation of the NPMLE in the proportional hazards mixed-effects model. *Statistica Sinica* 2009; **19**:997–1011.

Vaida F, Blanchard S. Conditional Akaike information for mixed-effects models. *Biometrika* 2005; **92**:351–370.

Donohue MC, Overholser R, Xu R, Vaida F. Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika* 2011; **98**(3):685–700. URL http://biomet.oxfordjournals.org/cgi/content/abstract/98/3/685.

Zeng D, Lin DY. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Series B* 2007; **69**:507–564.

Liu I, Blacker DL, Xu R, Fitzmaurice G, Tsuang MT, Lyons MJ. Genetic and environmental contributions to age of onset of alcohol dependence symptoms in male twins. *Addiction* 2004b; **99**:1403–1409.

Brown GG, Rutemiller HC. Means and variances of stochastic vector products with applications to random linear models. *Management Science* 1977; **24**(2):210–216.

Schoenfeld DA. Partial residuals for the proportional hazards regression model. *Biometrika* 1982; **69**:239–241.

O'Quigley J, Flandre P. Predictive capability of proportional hazards regression. *Proc. of the National Academy of Science USA* 1994; **91**:2310–2314.

Huang J, Harrington D. Penalized partial likelihood regression for right-censored data with bootstrap selection of the penalty parameter. *Biometrics* 2002; **58**:781–791.

Kent JT. Information gain and a general measure of correlation. *Biometrika* 1983; **70**:163–174.

Kent JT, O'Quigley J. Measures of dependence for censored survival data. *Biometrika* 1988; **75**:525–534.

Xu R, O'Quigley J. A R^2 measure of dependence for proportional hazards models. *Nonparametric Statistics* 1999; **12**:83–107.

O'Quigley J, Xu R, Stare J. Explained randomness in proportional hazards models. *Statistics in Medicine* 2005; **24**:479–489.

Cox DR, Snell EJ. *The Analysis of Binary Data (2nd ed.)*. Chapman and Hall, 1989.

Gray R. Tests for variation over groups in survival data. *Journal of the American Statistical Association* 1995; **90**:198–203.

Xu R, Vaida F, Harrington DP. Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Statistica Sinica* 2009; **19**:819–842.

Dunson DB, Chen Z. Selecting factors predictive of heterogeneity in multivariate event time data. *Biometrics* 2004; **60**:352–358.

Gray R. Correspondence (Re: Dunson and Chen, 2004). *Biometrics* 2006; **62**:623–624.

Lee KE, Kim Y, Xu R. Bayesian variable selection under the proportional hazards mixed-effects model. *Computational Statistics and Data Analysis* 2014; **75**:53–65.

Some Misconceptions on the Use of Composite Endpoints

Jianjun (David) Li and Jin Xu

Abstract Composite endpoint has been used frequently as a primary endpoint in clinical trials. However, there are some misconceptions on the use of composite endpoints. This paper identifies these misconceptions and discusses how to avoid them.

Keywords Composite endpoint • Component endpoint • Study design • Clinical trial

1 Introduction

Composite endpoint is frequently used as a primary endpoint in clinical trials. There are number of reasons to use a composite endpoint (Sankoh et al. 2014). Two basic reasons are: (1) No single endpoint can serve as the primary endpoint to demonstrate the drug effect with a reasonable sample size; (2) The drug is likely to have a similar magnitude of effect on several clinically meaningful endpoints. For these reasons, the clinically meaningful endpoints can be combined to form a composite endpoint as the primary endpoint. Each individual endpoint is then called component endpoint. A trial published recently at the New England of Journal of Medicine illustrates the use of composite endpoints. The ELIXA trial was a randomized and placebo-controlled trial to assess the effects of lixisenatide on cardiovascular morbidity and mortality among patients with type 2 diabetes (Pfeffer et al. 2015). The primary endpoint was a composite endpoint with the following components: death from cardiovascular causes, nonfatal myocardial infarction, nonfatal stroke, and hospitalization for unstable angina. The reason of using the composite endpoint was not stated in the paper but is clear from the event rates of individual components in the trial. According to the paper, the event rate of the

J.D. Li (✉)
Pfizer, Inc., Collegeville, PA 19426, USA
e-mail: david.li1@pfizer.com

J. Xu
Merck Sharp & Dohme, Kenilworth, New Jersey, USA

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_15

composite endpoint in placebo group was 13.2 % while the event rate of death from cardiovascular causes was just 3.1 %. Using death from cardiovascular causes as the primary endpoint will likely require a larger sample size as its rate is about 1/5 of the rate of the composite endpoint. It makes sense to use a composite endpoint to make the design more efficient—smaller and quicker.

The use of a composite endpoint is usually justified if the following assumptions are respected:

– The individual components of the composite are clinically meaningful and of similar importance to the patient.
– The expected effect on each component is similar based on biological plausibility.
– The clinically more important components of the composite should at least not be affected negatively.

These assumptions have been reflected in the regulatory guidelines. For example, the EMA guideline "Points to consider on multiplicity issues in clinical trials" (EMA) states that

> "When defining a composite variable it is recommended to include only components for which it can be assumed that treatment will influence them similarly."

The US FDA's guidance "Clinical studies section of labeling for human prescription drug and biological products – content and format" (FDA) specifically recommends that

> "In general, the results for all components of a composite endpoint should be presented. Presentation of all components reveals which components are driving the result and which components may be unaffected, or even adversely affected, by treatment with the drug."

The assumptions are seemingly easy to comprehend. However, they can be misunderstood sometimes, which leads to the misuse of composite endpoints. Many people believe that the results from individual components of a composite endpoint should be provided. However, if not done appropriately, this practice may result in misleading conclusions. An interesting example is shown in Table 1 (provided by Lubsen and Kirwan 2002). Suppose that the rates in Table 1 represent the results of a trial comparing a testing drug with a placebo. The composite endpoint was total mortality, 20 % for both groups and the rates of hospitalization were different, 25 % in the drug group and 35 % in the placebo group. One might mistakenly interpret the results by claiming that the drug had no impact on mortality but reduced hospitalization. This shows that examining the drug effect by looking at

Table 1 Death and hospitalization in a hypothetical trial

	Drug (%)	Placebo (%)
Died never hospitalized	15	5
Hospitalized and then died	5	15
Hospitalized and alive	20	20
None of above	60	60

the results from individual components separately could be misleading. Lubsen and Kirwan have warned that "distortion may occur when a specific clinical event, such as hospitalization, is analyzed while ignoring circumstances (such as death) that preclude the occurrence of the event considered".

This example reveals the complexity of using composite endpoints. A marginal analysis of each component may not help understand composite endpoints. The result of a soft component (e.g. hospitalization) may lead to a wrong conclusion if the result of hard component (e.g. death) is ignored. By extending the observation by Lubsen and Kirwan, this paper further shows that contrary to the traditional belief, reporting results of all components marginally could mask the contributions of individual components and in worse cases may lead to a wrong conclusion about the contributions of components.

Second common belief on composite endpoints is that if the testing drug has a similar effect on a few endpoints, combining these endpoints to form a composite endpoint will result in a higher event rate and increase statistical precision that improves the power of study. This paper will show that this is not always the case.

After pointing out these two misconceptions in Sect. 2, this paper will discuss when a composite endpoint should be used as the primary endpoint in Sect. 3 and how to design and analyze trials with a composite endpoint as the primary endpoint in Sect. 4, and provide summary and remarks in Sect. 5.

2 Two Misconceptions

2.1 Misconception on Use of Composite Endpoint

Suppose that a drug is effective in reducing mortality rate by 50 % (placebo rate 80 % vs drug rate 40 %). Also assume that the drug is effective in reducing hospitalization: Among those who are alive, the chance of hospitalization is also reduced by 50 % (placebo rate 20 % vs drug rate 10 %). Given these information, we can ascertain that the drug is effective in both mortality and hospitalization. So should we use the composite endpoint of death or hospitalization, instead of the endpoint of death, as the primary endpoint for demonstrating the drug efficacy?

Table 2 below provides a possible outcome of the trial. Note that 6 % of patients who survive are hospitalized in the drug group and 4 % in the placebo group, that is, 60 % of patients in the drug group survive and 10 % of these survivors are hospitalized whereas 20 % of patients in the placebo group survive and 20 % of them are hospitalized.

For the endpoint of death, the treatment difference is −40 %, but for the composite endpoint of death or hospitalization, the treatment difference is reduced to −38 %. Adding the soft endpoint of hospitalization to the hard endpoint of death in this case actually reduces the magnitude of treatment difference and consequently is not likely to improve the study power. Note that the treatment difference is similar for both the endpoint of death and the endpoint of hospitalization.

Table 2 Hypothetical outcome of a trial

	Drug (%)	Placebo (%)	Difference (%)
Died never hospitalized	1	2	−1
Hospitalized and then died	39	78	−39
Hospitalized and alive	6 (10% * 60%)	4 (20% * 20%)	+2
Component: death	40	80	−40
Component: hospitalization	45	82	−37
Composite: death or hospitalization	46	84	−38

Table 3 Death and hospitalization in a hypothetical trial

	Drug (%)	Placebo (%)	Difference (%)
Died never hospitalized	1	2	−1
Hospitalized and then died	39	78	−39
Hospitalized and alive	4 (6.7% * 60%)	4 (20% * 20%)	0
Component: death	40	80	−40
Component: hospitalization	43	82	−39
Composite: death + hospitalization	44	84	−40

From Table 2, we can see that even if all components have similar effects a composite endpoint may still not necessarily be a better choice. We need to consider the net contribution by a soft endpoint if we want to include it to form a composite endpoint.

When is the composite endpoint better than the hard endpoint? Per Table 3, in order to have non-negative contribution by hospitalization endpoint, the drug group should have a hospitalization rate of 6.7 % among survivors. The risk of hospitalization among survivors should be reduced by 67 % (placebo rate 20 % vs drug rate 6.7 %) instead of 50 %.

In practice, composite endpoints are often used for the purpose of increasing study power to have an efficient design. However, if the net effect by an additional soft endpoint is not carefully assessed, the inclusion of the soft endpoint may actually reduce the study power. The PROactive trial (Dormandy et al. 2005) is a good example. The CAPRICORN trial (The CAPRICORN Investigators 2001) discussed in Sect. 2.2 is another example.

2.2 *Misconception on Interpretation of Composite Endpoint*

Let us use the CAPRICORN trial as an example to show how the results of a composite endpoint and relevant component endpoints can be misinterpreted if they are not properly presented. The CAPRICORN trial was a randomized controlled trial to compare carvedilol and placebo. Two primary endpoints were pre-specified: (1) All-cause mortality and (2) all-cause mortality or cardiovascular hospital admission.

Table 4 Results from CAPRICORN trial

Primary endpoint	Carvedilol (N = 975)	Placebo (N = 984)
All-cause mortality n (%)	116 (12 %)	151 (15 %)
All-cause mortality or cardiovascular-cause hospital admission n (%)	340 (35 %)	367 (37 %)

Table 5 Results from CAPRICORN trial

Primary endpoint	Carvedilol (N = 975)	Placebo (N = 984)
All-cause mortality n (%)	116 (12 %)	151 (15 %)
Cardiovascular-cause hospital admission *among those survived* n (%)	224 (26 %)	216 (26 %)

Table 6 Hypothetical results from CAPRICORN trial

Primary endpoint	Carvedilol (N = 975 * 20)	Placebo (N = 984 * 20)
All-cause mortality n (%)	0 (0 %)	151 * 20 (15 %)
Cardiovascular-cause hospital admission n (%)	224 * 20 (23 %)	216 * 20 (22 %)
All-cause mortality or cardiovascular-cause hospital admission n (%)	224 * 20 (23 %)	367 * 20 (37 %)
Cardiovascular-cause hospital admission *among those survived* n (%)	224 * 20 (23 %)	216 * 20 (26 %)

Table 4 below, which is adopted from Table 2 in (The CAPRICORN Investigators 2001), summarizes the results of both primary endpoints. The marginal summary data for the endpoint of cardiovascular-cause hospital admission are not available from the reference. With the inclusion of the endpoint of cardiovascular-cause hospital admission, the difference between carvedilol and placebo was reduced from 3 % (=15–12 %) to 2 % (=37–35 %), so the inclusion actually led to a smaller treatment difference. Then what can we say about the drug effect on the endpoint of cardiovascular-cause hospital admission? Unaffected or adversely affected?

Given the negative impact on the treatment difference by inclusion of the endpoint of cardiovascular-cause hospital admission, the possible answer is: adversely affected. But let us present the results of CAPRICORN trial in a different way as in Table 5. In the carvedilol group, 116 patients died so 859 patients survived. The rate of cardiovascular-cause hospital admission among those who survived in the carvedilol group was 26 % (=224/859). In the placebo group, 833 patients survived and the rate was also 26 % (=216/833). So carvedilol did not increase the rate of cardiovascular-cause hospital admission among those who survived.

Let us take a look at an extreme hypothetical example where carvedilol was 100 % efficacious in reducing mortality so there was no death in the carvediol group and all death events were in the placebo group, and there was no overlap in events, i.e., no subject was hospitalized before he/she died. We also increase the N to 20 times larger as shown in Table 6 so the numerical differences could be statistically significant.

Based on the marginal result for the endpoint of cardiovascular-cause hospital admission, there was a high percentage of hospital admission in the carvedilol group (23 % vs. 22 %), with p-value <0.01. We might conclude that carvedilol prevented death but likely resulted in more hospitalizations. But looking at the result for cardiovascular-cause hospital admission among those who survived, the rate was actually lower in the carvedilol group, with p-value <0.001. So carvedilol not only prevented death but also reduced the chance of hospital admission among those who survived.

In summary, presenting results for individual components marginally could be misleading. A more informative presentation of the trial results is to present the result for the hard endpoint and to present the conditional result for the soft endpoint, to show the net contribution of the drug by the soft endpoint.

3 When to Use a Composite Endpoint

From Sect. 2.1, we see that it is not always best to use a composite endpoint and in order to make a composite endpoint a better choice, the effect of drug on an additional endpoint needs to be quite large. This section looks into when to use a composite endpoint or under what condition we should use a composite endpoint.

Table 7 represents the results of death and hospitalization in a clinical trial. The rate of hospitalization among survivors is calculated as the number of hospitalization events divided by the number of survivors. It can follow that the rate of composite endpoint in the drug group can be parsed into two parts:

$$r_{C,d} = \frac{n_{D,d} + n_{aH,d}}{N_d} = \frac{n_{D,d} + r_{aH,d}\left(N_d - n_{D,d}\right)}{N_d} = r_{D,d} + r_{aH,d}\left(1 - r_{D,d}\right).$$

Similarly, $r_{C,p} = r_{D,p} + r_{aH,p}\left(1 - r_{D,p}\right)$. So the treatment difference in the composite endpoint is

$$r_{C,d} - r_{C,p} = r_{D,d} - r_{D,p} + \left[r_{aH,d}\left(1 - r_{D,d}\right) - r_{aH,p}\left(1 - r_{D,p}\right)\right]$$

A negative value of $r_{C,d} - r_{C,p}$ favors the drug group. To have a larger treatment effect in favor of drug in the composite endpoint than in the endpoint of death, we

Table 7 Death and hospitalization in a trial

	Drug	Placebo	Difference
Death	$r_{D,d}\left(= \frac{n_{D,d}}{N_d}\right)$	$r_{D,p}\left(= \frac{n_{D,p}}{N_p}\right)$	$r_{D,d} - r_{D,p}$
Hospitalization among survivors	$r_{aH,d}\left(= \frac{n_{aH,d}}{N_d - n_{D,d}}\right)$	$r_{aH,p}\left(= \frac{n_{aH,p}}{N_p - n_{D,p}}\right)$	
Composite: death or hospitalization	$r_{C,d}\left(= \frac{n_{D,d} + n_{aH,d}}{N_d}\right)$	$r_{C,p}\left(= \frac{n_{D,p} + n_{aH,p}}{N_p}\right)$	$r_{C,d} - r_{C,p}$

need $r_{aH,d}\left(1-r_{D,d}\right)-r_{aH,p}\left(1-r_{D,p}\right)<0$. That is

$$rr_{aH}=\frac{r_{aH,d}}{r_{aH,p}}<\frac{1-r_{D,p}}{1-r_{D,d}}=\frac{1-r_{D,p}}{1-rr_D\cdot r_{D,p}}$$

where $rr_{aH}\left(=\frac{r_{aH,d}}{r_{aH,p}}\right)$ is the observed relative risk of hospitalization among survivors (drug vs placebo) and $rr_D\left(=\frac{r_{D,d}}{r_{D,p}}\right)$ is the observed relative risk of death (drug vs placebo). We consider the case $rr_D<1$ because $rr_D\geq 1$ would indicate that the risk of death in the drug group is higher than or equal to that in the placebo group. If $rr_D<1$, $\frac{1-r_{D,p}}{1-rr_D\cdot r_{D,p}}<1$. Obviously $rr_{aH}<1$ does not automatically guarantee $rr_{aH}<\frac{1-r_{D,p}}{1-rr_D\cdot r_{D,p}}$. So even if the drug reduces the risk of hospitalization among survivors, the composite endpoint may not necessarily yield a larger treatment effect. If $rr_D=0$, $\frac{1-r_{D,p}}{1-rr_D\cdot r_{D,p}}=1-r_{D,p}$, which is the smallest possible value for $\frac{1-r_{D,p}}{1-rr_D\cdot r_{D,p}}$. So if the drug is very effective on the endpoint of death so that the observed relative risk of death (rr_D) is very close to 0, caution should be exercised in using the composite endpoint. This is true especially when the background rate of death ($r_{D,p}$) is not small. When rr_D is not very small but $r_{D,p}$ is very small, $\frac{1-r_{D,p}}{1-rr_D\cdot r_{D,p}}$ is close to 1. It is more likely that $rr_{aH}<\frac{1-r_{D,p}}{1-rr_D\cdot r_{D,p}}$, which would yield a larger treatment effect if the composite endpoint is used.

4 Analysis Approach

From Sect. 3, we see that composite endpoints may not always be the best choice. If at the clinical trial design stage we are not clear whether we should use a composite endpoint, one thing we can do is to designate both the hard endpoint death and the composite endpoint as primary endpoints and apply a multiplicity control procedure to control the familywise type I error rate. For example, we can use the following testing procedure: Claim that the drug is effective if

$$p_C\leq 0.025 \text{ and } rr_D<1 \text{ and } rr_{aH}<\frac{1-r_{D,p}}{1-rr_D\cdot r_{D,p}} \quad \text{or}$$

$$p_D\leq\alpha^* \text{ and } rr_D<1 \text{ and } \frac{1-r_{D,p}}{1-rr_D\cdot r_{D,p}}\leq rr_{aH}<1$$

where p_C is one-sided p-value for the composite endpoint and p_D is one-sided p-value for the hard endpoint death, and $\alpha*$ is calculated to satisfy the following probability inequality

$$\Pr\Big(p_C \le 0.025,\ rr_D < 1,\ \text{and}\ rr_{aH} < \tfrac{1-r_{D,p}}{1-rr_D\cdot r_{D,p}}\ \text{ or }\ p_D \le \alpha^*,\ rr_D < 1,\ \text{and}\ \tfrac{1-r_{D,p}}{1-rr_D\cdot r_{D,p}} \le rr_{aH} < 1\Big) \le 0.025$$

which is evaluated under the null hypothesis that the drug is the same as placebo, i.e., $t_{D,d} = t_{D,p}$ and $t_{C,d} = t_{C,p}$, for all possible values of $t_{D,d}$ and $t_{C,d}$, where $t_{D,d}$, $t_{D,p}$ are the true rate of death in the drug and placebo group, respectively, and $t_{C,d}$, $t_{C,p}$ are the true rate of composite endpoint in the drug and placebo group, respectively.

5 Summary and Discussions

Composite endpoint has been used quite frequently as a primary endpoint in clinical trials. However, as illustrated in this paper the use of composite endpoints does not necessarily achieve the desirable outcome. In many trials, using of a composite endpoint actually results in a larger p-value compared with using a hard endpoint alone. This paper shows some likely causes of this phenomenon. Contrary to common perceptions, a large treatment effect is not a guaranteed outcome from using a composite endpoint even if the drug has a similar effect on all components. In order for a composite endpoint to be a favorable choice over its component hard endpoint, i.e., to observe a larger effect on the composite endpoint, the net contribution to the composite endpoint by the soft component has to be decently large.

Currently a common practice in reporting composite endpoints is to report each component endpoint marginally. Using an example, we show that this type of reporting may lead to wrong conclusions. A positive net effect by drug on the soft endpoint could be masked and can be misread as negative effect in certain situations. So we suggest reporting the marginal result for the hard endpoint and conditional result for the soft endpoint to make the net contribution of soft endpoint transparent, which helps assess whether the drug has additional effect on the soft endpoint.

At the clinical trial design stage, it is challenging sometimes to know whether a composite endpoint has a larger effect than the hard endpoint. So it may be advisable to designate both the composite endpoint and the hard endpoint as primary endpoints and use an appropriate statistical method to control the familywise type I error rate. We proposed one method in this paper. Further research is needed to investigate whether there is a better alternative method.

References

Sankoh, A.J., Li, H., D'Agostino, R.B. Use of composite endpoints in clinical trials. Statist. Med. 2014, 33 4709–4714.

Pfeffer, M.A., Claggett, B., Diaz, R., et al. Lixisenatide in Patients with Type 2 Diabetes and Acute Coronary Syndrome. N Engl J Med 2015; 373:2247-2257.

EMA. Points to consider on multiplicity issues in clinical trials. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf.

FDA. Clinical studies section of labeling for human prescription drug and biological products - content and format. http://www.fda.gov/RegulatoryInformation/Guidances/ucm127509.htm.

Lubsen, J. and Kirwan, B. (2002). Combined endpoints: can we use them? Statist. Med. 2002; 21:2959–2970.

Dormandy, J. A., Charbonnel, B., Eckland, D.J., et al. Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive Study (PROspective pioglitAzone Clinical Trial In macroVascular Events): a randomised controlled trial. *Lancet* **366**, 1279-1289.

The CAPRICORN Investigators. Effect of carvedilol on outcome after myocardial infarction in patients with left-ventricular dysfunction: the CAPRICORN randomised trial. *Lancet* 357, No. 9266, p1385–1390.

Part V
Clinical and Safety Monitoring in Clinical Trials

A Statistical Model for Risk-Based Monitoring of Clinical Trials

Gregory J. Hather

Abstract Risk-based monitoring allows monitors of clinical trial sites to focus their visits on sites with the greatest potential for risk reduction. Here we present a statistical model that recommends sites for the monitor to visit. The model makes use of a pre-visit assessment supplied by the monitor, as well as other measurable factors, to predict the monitor's post-visit assessment of the risk reduction resulting from the visit. The monitor is then directed to visit the sites with the highest predicted risk reduction. We demonstrate the properties of this model using a simulation. Our simulation compares two strategies for directing monitors, one of which relies on the model, while the other strategy relies only on the monitor's pre-visit assessments. Our simulation demonstrates that the model-based strategy can direct the monitors to sites with greater potential for risk reduction. Finally, we discuss alternative models as well as potential pitfalls of risk-based monitoring.

Keywords Clinical trial • Oversight • Quality • Risk based monitoring

1 Background/Approach

Onsite monitoring of clinical trial sites is a common means to ensure protection of the rights and safety of human subjects and the quality of the data (Williams 2006; Morrison et al. 2011; Bhatt 2011). Monitors may identify problems such as data entry errors, protocol deviations, or inadequate staff training at the site. However, for a large pharmaceutical company with thousands of ongoing clinical trial sites, monitoring is a significant expense. In recent years, some companies have implemented risk-based monitoring (TransCelerate BioPharma Inc 2013), where inspectors are sent to the sites predicted to have high risk. This strategy can potentially save money through more efficient use of monitoring resources. Risk-based monitoring is in contrast to more traditional methods of directing monitors, such as relying on regular inspection schedules.

G.J. Hather (✉)
Takeda Pharmaceuticals, 35 Landsdowne Street, Cambridge, MA 02139, USA
e-mail: Greg.Hather@takeda.com

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_16

The use of risk-based monitoring has been facilitated by the introduction of systems for central monitoring (U.S. Department of HHS, FDA 2013). Central monitoring is the practice of remotely collecting, visualizing, and analyzing data from the all the clinical trial sites. The analysis of such data can both identify problems at a site and potential risk factors. Thus central monitoring increases monitoring effectiveness in two ways: by identifying problems remotely and by shifting onsite visits from low risk sites to high risk sites. However, central monitoring has not at present eliminated the use of onsite visits (Hullsiek et al. 2015), and many companies use a mix of both.

Risk-based monitoring for clinical trial sites is a relatively recent development, and several papers have encouraged the use of this method (Franco et al. 2013; Burgess and ICONIK 2013). Regulatory agencies have also issued guidance on the use of risk-based monitoring (U.S. Department of HHS, FDA 2013; European Medicines Agency 2013). However, the field is still at an early stage, and the best methods for risk-based monitoring are not yet known.

The most commonly used statistical methods for risk-based monitoring involve performing statistical tests to identify sites that are in some way different from the other sites (Timmermans et al. 2015, 2016; Desmet et al. 2014; Venet et al. 2012). Other methods focus on data integration and visualization of measured risk indicators, with less emphasis on statistical modeling (Zink 2014; Taylor et al. 2002). Some researchers have developed statistical methods to specifically identify sites where data fraud is occurring (Taylor et al. 2002; Buyse et al. 1999; Al-Marzouki et al. 2005; George and Buyse 2015). Other researchers have considered risk-based methods specifically focusing on source data verification (Tantsyura et al. 2015; Nielsen et al. 2013; van den Bor et al. 2016).

While risk indicators are often used to suggest sites to visit, in practice, monitors often combine the numerical risk ranking with their own knowledge of the protocol and trial sites that may not be easily captured by the model. For example, the monitor may know that a certain protocol is harder to implement than the other study protocols, but such a factor may be neglected in a quantitative risk-based model. Likewise, the monitor may observe at an earlier site visit that the investigator was disorganized. This observation may not be easily integrated into conventional models for risk-based monitoring. To the knowledge of the author, no statistical model for directing inspectors of clinical trial sites has yet been published that explicitly combines both the monitors' judgements with other quantitative risk factors.

Ideally, monitors should be directed to sites that are highly worthwhile to visit because the sites allow for the largest reduction in risk relative to the cost and time required to visit the site. Here, we present a model that predicts which sites will be worthwhile to visit. Our approach uses subjective assessments, both pre and post-visit, of the worthwhileness of a visit. Measured factors, along with the pre-visit assessment, can be used to predict the post-visit assessment. The monitor can then inspect sites that are predicted to be the most worthwhile to visit.

2 Methods

2.1 Model

We assume that at regular intervals, each monitor remotely reviews all the sites in his or her territory. The monitor then takes a pre-visit survey regarding how worthwhile it is to visit each site. If a site is selected for an onsite visit, then the monitor also performs a post-visit assessment about how worthwhile the visit was.

For concreteness, we assume the remote reviews are performed monthly. At the beginning of the month, the monitor would take a survey for each site by responding to the following statement: "It is worthwhile to visit this site this month". The possible responses would be: "Strongly agree", "Agree", "Neutral", "Disagree", or "Strongly disagree" (DeMars 2010). The monitor would also take a similar survey after visiting a site (a post-visit assessment), by responding to the statement: "It was worthwhile to visit this site this month".

Initially, the sites to visit would be selected by ranking the sites by the pre-visit assessment and breaking ties randomly. However, once sufficient data was collected, we would fit a regression model that predicts the post-visit assessments from the pre-visit assessments and other factors specific to each clinical trial site. There are many potential predictive factors. Here, we consider the country where the site is located, as well as the protocol, phase, and therapy area for the study. We propose a mixed-effects model (Crawley 2012) to predict the post-visit assessment of site i and time j

$$\begin{aligned}(\text{post-visit assessment})_{\mathrm{ij}} = {} & c^{*}(\text{pre-visit assessment})_{\mathrm{ij}} + \text{intercept} \\ & + (\text{country effect})_{\mathrm{i}} + (\text{therapy area effect})_{\mathrm{i}} \\ & + (\text{phase effect})_{\mathrm{i}} + (\text{protocol effect})_{\mathrm{i}} + \text{error}_{\mathrm{ij}}\end{aligned}$$

Here, the first two terms are fixed effects, and c is a coefficient. The country, therapy area, phase, and protocol effects are random effects. The pre and post-visit assessments would be converted into numbers (5 for "Strongly agree" and 1 for "Strongly disagree").

The random effects are assumed to be normally distributed and independent of each other. The error term is assumed to be normally distributed and independent for each visit. The random effects and the error term were assumed to have a mean of zero. The parameters are assumed to be constant over time. If a particular factor level (e.g. a particular country) was present in the test dataset but not in the training dataset, then the coefficient for that factor would be set to zero when making the prediction in the training set.

Given that the error terms are independent, there will not be temporal correlation of the post-visit assessments at any given site unless there is temporal correlation in the pre-visit assessments. However, if we look across sites, sites with higher post-visit assessments in a given month will tend to have higher post-visit assessments in other months because the random effects terms for these sites are expected to be higher compared with other sites.

One advantage of using random effects terms in the model is that it prevents certain parameter estimates from becoming extreme when there is little data for a particular level of a factor (e.g. a particular country). This is because the random effects terms for a particular factor (e.g. country) are assumed to come from a common distribution. This assumption allows for the pooling of information across factor levels, which helps moderate the parameter estimates.

Although pre-visit assessments would be available for all the sites each month, only the sites that were visited, and thus had post-visit assessments, could be used to estimate the parameters. After estimating the model parameters, we would apply the fitted model to predict post-visit assessments for all the sites every month following the model fitting. The monitor would then visit sites predicted to have the highest (most worthwhile) post-visit assessments.

2.2 Simulation

To the author's knowledge, there are currently no publicly available datasets to which the model could be applied. Therefore, we created a simulated dataset to demonstrate our model. For this dataset, we assumed there were 1000 sites, 10 monitors, and 20 countries. We assumed that each monitor could visit 10 sites every month, and that 20 months of data were available. We simulated covariates, model parameters, and post-visit assessments.

We assumed that each site was randomly assigned to a country with equal probability. We assumed that each monitor oversaw the sites in two countries. We also assumed that there were 50 protocols, and each site was assigned to a protocol with equal probability. We assumed that each protocol was assigned to one of three phases with equal probability and one of five therapy areas with equal probability. Finally, we assumed that the pre-visit assessment at each month was drawn from the following distribution with no temporal correlation: 5 % "Strongly disagree", 25 % "Disagree", 40 % "Neutral", 25 % "Agree", and 5 % "Strongly agree".

For the model, we generated country effects, protocol effects, phase effects, and therapeutic area effects by sampling from a normal distribution with mean 0 and a standard deviation of 0.25. We assumed that the intercept was 0.1, the pre-visit coefficient was 0.9, and the error term had a standard deviation of 0.5. We assumed that the factors for the pre-visit assessment were 1, 2, 3, 4, and 5, respectively, for "Strongly disagree", "Disagree", "Neutral", "Agree", and "Strongly agree", respectively. We used the simulated parameters and covariates to generate the post-visit assessments. Finally, we rounded each simulated post-visit assessment to the nearest integer between 1 and 5 to ensure that it corresponded to one of the five possible responses.

In order to provide an unbiased assessment of our model's performance, we split the data into a training set and a test set. The training dataset was used to estimate the model parameters, while the test dataset was used to evaluate different monitoring strategies. We defined the first 10 months as the training set and the last 10 months as the test dataset.

2.3 Model Fitting

We used the lme package (Crawley 2012) in R to fit the training data. The training data consisted of 1000 rows (10 monitors visiting 10 sites each month for 10 months).

2.4 Evaluating Monitoring Strategies

We decided to compare two different monitoring strategies applied to the test data. The first strategy we called "intuition-based" because the monitor's subjective pre-visit assessment was the only factor considered in assigning the visits. This strategy simply ranks the sites by the pre-visit assessment, randomly breaks ties, and then assigns the monitor to visit the top ten sites.

The second strategy we called "model-based" because the model was used to direct the monitor. This strategy ranks the sites by the predicted post-visit assessment, randomly breaks ties, and then assigns the monitor to visit the top ten sites. Note that this strategy makes use of both the monitor's subjective pre-visit assessment and other measurable factors.

These two strategies were applied to the same test set, and the average observed post-visit assessment for the sites selected by each strategy was compared. Note that the averages were expected to be different because the strategies would select different sites to visit. In order to determine which strategy tended to produce higher averages, the entire simulation was repeated 100 times. Our code is available at https://goo.gl/dOA5H3.

3 Results

First, we investigated the pre-visit assessments in the training dataset pooled over the 100 simulations. Figure 1a shows the distribution of the assessments for all sites, while Fig. 1b shows the distribution of the assessments for only the visited sites. As expected, sites with low pre-visit assessments were not visited.

Next, we investigated the post-visit assessments in the test dataset pooled over the 100 simulations. Figure 2a shows the distribution of the post-visit assessment

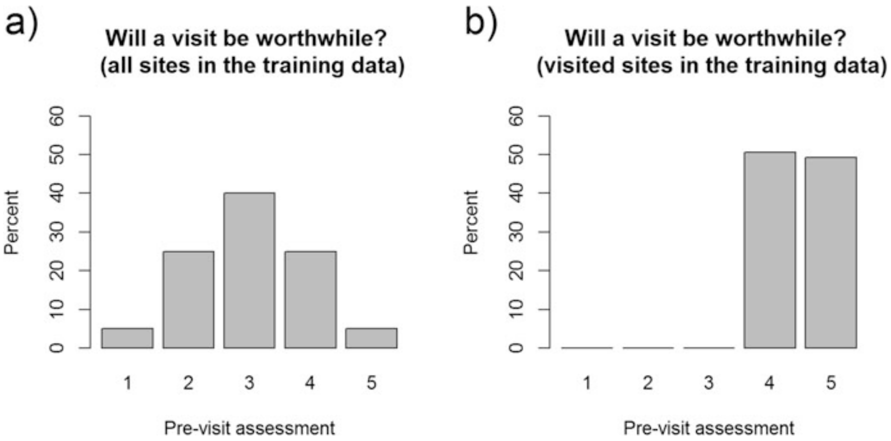


Fig. 1 Pre-visit assessments in the training data. (**a**) The distribution of the pre-visit assessments for all sites. (**b**) The pre-visit assessment distribution for the visited sites only

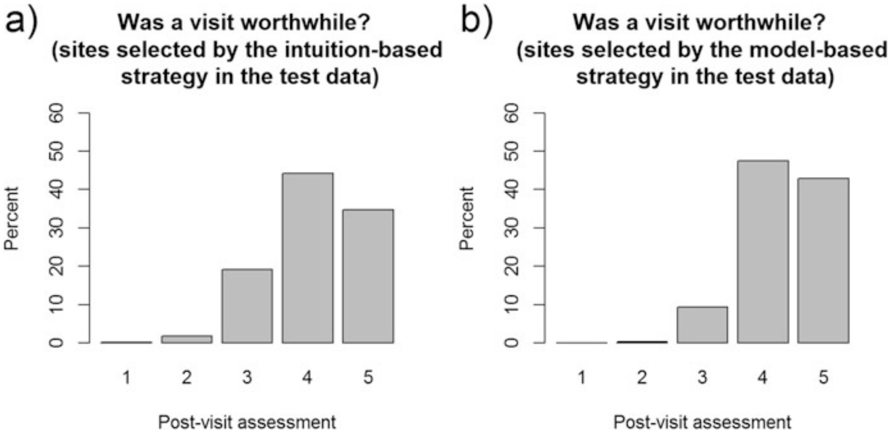


Fig. 2 Post-visit assessments in the testing data. (**a**) The post-visit assessment distribution visited sites using the intuition-based strategy. (**b**) The same result for the model-based strategy

for only the visited sites when the intuition-based strategy was used, while Fig. 2b shows the same result for the model-based strategy. The model-based strategy resulted in a higher mean for the post-visit assessments (4.33 vs 4.12, Wilcoxon signed rank test p-value $= 4 \times 10^{-18}$). As another way of measuring the performance, we considered visits with a post-visit assessment of "Neutral" or below to be of low value. We found the model-based strategy resulted in a lower portion of low value visits (9.6 % vs 21.0 %, Wilcoxon signed rank test p-value $= 4 \times 10^{-18}$).

4 Conclusions

Our results demonstrate that model-based direction of monitors has the potential to allow more efficient use of monitoring resources. Without real data, though, we cannot know how large this increase in efficiency would be. If applied to real data, we expect to find a different level of savings, which may or may not be large enough to justify the cost of the data collection and analysis. Therefore, the result should be considered proof of principle, rather than a recommendation that model-based direction be used in all cases.

In our simulation, the pre-visit assessment was an essential covariate in the model. Note that if this term had not been used, the model could potentially be less accurate than educated guesses made by the inspectors. Using such a model could direct monitors to sites that were on average less worthwhile to visit. In the other extreme, one could imagine a scenario where the pre-visit assessments add very little predictive power to the model. In this case, one could eliminate the pre-visit assessments to save the monitors' time.

There are many potential improvements to the model. For example, additional covariates could be added. These covariates could include information about the principal investigator, the current enrollment, the monitor's past assessments, or the time since the last visit. One could also consider using a Hidden Markov Model (Rabiner and Juang 1986), as this type of model may better describe the evolution of the sites' risk over time. Our approach could also be modified so that the parameters estimates are updated continuously as new data is acquired.

We acknowledge several limitations of our approach. First, classifying the visit assessments into discrete groups may result in a loss of information, and thus the selection of less worthwhile sites. A second shortcoming is that the relationship between the site characteristics and the post-visit assessments may change over time, and this is not accounted for by our model. Finally, the subjective nature of the post-visit assessments make the benefit of improved site selection hard to quantify.

We note that the data used to fit the model should come from site visits that were selected solely based on measured variables. Otherwise, the data would be missing not at random (Rubin 1976), which may cause severe bias in the model fitting.

The topic of risk-based monitoring of clinical trials has potential connections with several other fields. For example, this work is related to decision support systems (Power et al. 2015), since our model helps monitors decide which site to visit. We also note that risk-based monitoring has been applied in other fields, such as engineering, where it has been used for several decades (Khan and Haddara 2003). In addition, decisions in consumer lending are often made by combining computer and human judgement about risk (Thomas 2000). Insights from these fields may help with risk-based monitoring for clinical trials.

In summary, we developed a model to predict the worthwhileness of a site visit by a monitor. This model was able to prescribe site visits that were expected to be more worthwhile. We demonstrated a simulated scenario where use of the model allowed monitors to select sites that were more worthwhile to visit compared with sites selected by the monitor's intuition alone. Our work demonstrates the potential of such a model to improve the choice of visited sites and reduce risk with limited resources.

Acknowledgments The author would like to thank Daniela Conrado, Beth Soffer, Shaum Kabadi, Phillip Wallis, and Aditya Rajan for their helpful discussions.

References

Williams, G.W., 2006. The other side of clinical trial monitoring; assuring data quality and procedural adherence. Clinical Trials, 3(6), pp.530-537.

Morrison, B.W., Cochran, C.J., White, J.G., Harley, J., Kleppinger, C.F., Liu, A., Mitchel, J.T., Nickerson, D.F., Zacharias, C.R., Kramer, J.M. and Neaton, J.D., 2011. Monitoring the quality of conduct of clinical trials: a survey of current practices. Clinical Trials, 8(3), pp.342-349.

Bhatt, A., 2011. Quality of clinical trials: A moving target. Perspectives in clinical research, 2(4), p.124.

TransCelerate BioPharma Inc. (2013). Position paper: risk-based monitoring methodology. http://www.transceleratebiopharmainc.com/wp-content/uploads/2013/10/TransCelerate-RBM-Position-Paper-FINAL-30MAY2013.pdf.

Oversight of Clinical Investigations: A Risk-Based Approach to Monitoring. U.S. Department of HHS, FDA, August 2013 OMB Control No. 0910–0733.

Hullsiek, K.H., Kagan, J.M., Engen, N., Grarup, J., Hudson, F., Denning, E.T., Carey, C., Courtney-Rodgers, D., Finley, E.B., Jansson, P.O. and Pearson, M.T., 2015. Investigating the Efficacy of Clinical Trial Monitoring Strategies Design and Implementation of the Cluster Randomized START Monitoring Substudy. Therapeutic innovation & regulatory science, 49(2), pp.225-233.

Franco, P., Hronec, M. and Karacsony, A., 2013. Risk-based monitoring: Reduce clinical trial costs while protecting safety and quality. PricewaterhouseCoopers.

Burgess, M. and ICONIK, I., 2013. Less is more: risk-based monitoring of site performance. ICON Insight, 13.

Reflection paper on risk based quality management in clinical trials. European Medicines Agency. 18 November 2013 EMA/269011/2013.

Timmermans, C., Doffagne, E., Venet, D., Desmet, L., Legrand, C., Burzykowski, T. and Buyse, M., 2016. Statistical monitoring of data quality and consistency in the Stomach Cancer Adjuvant Multi-institutional Trial Group Trial. Gastric Cancer, 19(1), pp.24-30.

Desmet, L., Venet, D., Doffagne, E., Timmermans, C., Burzykowski, T., Legrand, C. and Buyse, M., 2014. Linear mixed-effects models for central statistical monitoring of multicenter clinical trials. Statistics in medicine, 33(30), pp.5265-5279.

Venet, D., Doffagne, E., Burzykowski, T., Beckers, F., Tellier, Y., Genevois-Marlin, E., Becker, U., Bee, V., Wilson, V., Legrand, C. and Buyse, M., 2012. A statistical approach to central monitoring of data quality in clinical trials. Clinical Trials, 9(6), pp.705-713.

Timmermans, C., Venet, D. and Burzykowski, T., 2015. Data-driven risk identification in phase III clinical trials using central statistical monitoring. International journal of clinical oncology, pp.1-8.

Zink, R.C., 2014. Risk-based monitoring and fraud detection in clinical trials using JMP and SAS. SAS Institute.

Taylor, R.N., McEntegart, D.J. and Stillman, E.C., 2002. Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. Drug information journal, 36(1), pp.115-125.

Buyse, M., George, S.L., Evans, S., Geller, N.L., Ranstam, J., Scherrer, B., Lesaffre, E., Murray, G., Edler, L., Hutton, J. and Colton, T., 1999. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. Statistics in medicine, 18(24), pp.3435-3451.

Al-Marzouki, S., Evans, S., Marshall, T. and Roberts, I., 2005. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. Bmj, 331(7511), pp.267-270.

George, S.L. and Buyse, M., 2015. Data fraud in clinical trials. Clinical investigation, 5(2), pp.161-173.

Tantsyura, V., Dunn, I.M., Fendt, K., Kim, Y.J., Waters, J. and Mitchel, J., 2015. Risk-Based Monitoring A Closer Statistical Look at Source Document Verification, Queries, Study Size Effects, and Data Quality. Therapeutic Innovation & Regulatory Science, 49(6), pp.903-910.

Nielsen, E., Hyder, D. and Deng, C., 2013. A data-driven approach to risk-based source data verification. Therapeutic Innovation & Regulatory Science, p.2168479013496245.

van den Bor, R.M., Oosterman, B.J., Oostendorp, M.B., Grobbee, D.E. and Roes, K.C., 2016. Efficient Source Data Verification Using Statistical Acceptance Sampling A Simulation Study. Therapeutic Innovation & Regulatory Science, 50(1), pp.82-90.

DeMars, C. (2010). Item Response Theory. Oxford, UK: Oxford University Press.

Crawley, M. J. (2012) Mixed-Effects Models, in The R Book, Second Edition, John Wiley & Sons, Ltd, Chichester, UK.

Rabiner, L.R. and Juang, B.H., 1986. An introduction to hidden Markov models. ASSP Magazine, IEEE, 3(1), pp.4-16.

Rubin, D.B., 1976. Inference and missing data Biometrika 63 (3): 581–592. Find this article online.

Power, D.J., Sharda, R. and Burstein, F., 2015. Decision support systems. John Wiley & Sons, Ltd.

Khan, F.I. and Haddara, M.M., 2003. Risk-based maintenance (RBM): a quantitative approach for maintenance/inspection scheduling and planning. Journal of Loss Prevention in the Process Industries, 16(6), pp.561-573.

Thomas, L.C., 2000. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. International journal of forecasting, 16(2), pp.149-172.

Blinded Safety Signal Monitoring for the FDA IND Reporting Final Rule

Greg Ball and Patrick M. Schnell

Abstract We introduce a safety monitoring procedure for two-arm blinded clinical trials. This procedure incorporates a Bayesian hierarchical model for using prior information and pooled event rates to make inferences on the rate of adverse events of special interest in the test treatment arm. We describe a collaborative process for specifying the prior and calibrating the operating characteristics.

Keywords Adverse events • Bayesian inference • Blinded data • Clinical trials • Safety monitoring

1 Introduction

Regulatory requirements and guidance documents regarding clinical trials place primary responsibility on the sponsor for ongoing safety evaluation of investigational products (Guideline for Good Clinical Practice E6(R1) 1996; European Commission 2006; US Department of Health and Human Services Food and Drug Administration 2012), as the sponsor is best positioned to assess the overall safety of these drugs and devices. Sponsors should review aggregated safety data throughout the development program and facilitate early planning for assessment of emerging safety signals by establishing a multidisciplinary Safety Management Team (SMT) (US Department of Health and Human Services Food and Drug Administration 2012; Crowe et al. 2009; Xia et al. 2011; Chuang-Stein and Xia 2013). This collaborative team of subject matter experts from clinical, safety, and statistics groups could synthesize all available information to provide a complete assessment of the safety profile. Adverse events of special interest (AESI) could be established and analyses

G. Ball (✉)
Biostatistics and Research Decision Sciences, Merck Research Laboratories, 126 E Lincoln Ave, Rahway, NJ 07065-4607, USA
e-mail: greg.ball@merck.com

P.M. Schnell
Division of Biostatistics, School of Public Health, University of Minnesota, 420 Delaware St SE, Minneapolis, MN 55455-0381, USA

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_17

pre-specified in order to identify and understand potential safety signals as early as possible in the drug development process. The SMT would evaluate accumulating blinded data on a regular and ongoing basis and alert the Safety Review Committee if evidence of a higher than expected AESI rate were to emerge.

Regulations also assert that trials should only be carried out if the risks have been adequately assessed and can be appropriately managed (Guideline for Good Clinical Practice E6(R1) 1996; World Medical Association 2006; Council for International Organizations of Medical Sciences 2002). Potential issues that may be suspected because of preclinical data or other available sources should be targeted for evaluation (US Department of Health and Human Services Food and Drug Administration 2005). In the United States, sponsors must report a suspected adverse reaction if an aggregate analysis indicates that a specific event "occurs more frequently in the drug treatment group than in a concurrent or historical control group", including any clinically important increase over what is listed in the Investigator's Brochure (IB) or inferred from data in the IB (US Department of Health and Human Services Food and Drug Administration 2005, 2010). In the European Union, "an increase in the rate of occurrence . . . of an expected serious adverse reaction, which is judged to be clinically important" is a safety issue that requires expedited reporting (European Commission 2006). Sponsors and regulatory agencies, striving to protect patients while minimizing development costs, need innovative approaches that apply quantitative methods to the safety monitoring process while maintaining the blinding of interim data and ensuring trials lead to conclusive results.

Data Monitoring Committees (DMCs) are independent from the ongoing collection and analysis of trial data. They periodically evaluate accumulating unblinded data and make recommendations about the continuing safe conduct of trials. While all trials require safety monitoring, not all trials require a DMC, which adds administrative complexity and consumes additional resources (US Department of Health and Human Services Food and Drug Administration 2012). On the other hand, establishing a DMC does not diminish the sponsor's responsibility in the safety monitoring process. With or without a DMC, sponsors must evaluate accumulating data from ongoing trials, as well as other available information, to vigilantly watch over the evolving safety profile. Passive surveillance is a process for identifying disproportionately high rates of AE-drug combinations in large observational drug safety databases (Huang et al. 2014). This process does not allow for a comparison of relative risks in ongoing clinical trials. An active surveillance system is needed, which also enables continuous monitoring of a collection of prespecified AESI.

Few statistical methods have been developed directly for the purpose of active safety monitoring; however, methods developed for efficacy analyses can be adapted for the evaluation of safety data. One of the first sequential hypothesis tests, and for the purpose of testing two simple hypotheses the most efficient, is the sequential probability ratio test (SPRT) originally proposed by Wald (1945). Modifications to the SPRT have been proposed by Goldman (1987) and Goldman and Hannan (2001) to achieve specified operating characteristics for various statistical tests.

These sequential evaluations of the accumulating data provide a way to quickly and objectively identify safety concerns with minimum average sample size. Unfortunately, few trials conform to the restrictive design (two treatments, matched pairs, immediate outcomes, normal or binary data, and continuous monitoring), and regardless, the maximum sample size is unbounded (Pocock 1977). For these reasons, and since gains in the average sample size diminish with each successive look at the data, group sequential methods have been developed.

Unless there are big treatment differences, little benefit can be achieved with more than about five interim analyses (Pocock 1982). Group sequential methods by O'Brien and Fleming (1979) and Lan and DeMets (1983) are well known. In this frequentist paradigm, conclusions must depend on the stopping rule. However, if the likelihood principle can be accepted, then sequential analyses are not required and methods better adapted for informing decision-makers can be used (Anscombe 1963; Cornfield 1966, 1976). Safety monitoring in clinical trials is not confirmatory in nature; it is a learning process and reviews should be conducted with an analysis method designed for learning and making decisions. In addition, by the time Phase 2 studies are carried out, considerable information to form prior beliefs is often available from experience with earlier trials and related treatments. Bayesian methods can provide the same desirable properties as with sequential analyses, without the practical and conceptual difficulties (Herson 1979; Freedman and Spiegelhalter 1989; Freedman et al. 1994; Thall and Simon 1994; Thall et al. 1995; Heitjan 1997). Continuous monitoring of the data becomes possible with a more flexible approach and with greater ease of interpretation.

Bayesian hierarchical models, like Berry and Berry (2004) and Xia et al. (2011), offer the potential to borrow strength across subgroups in the data, such as with AE body systems. This is more helpful for signal detection of clusters of unknown AEs and less helpful for monitoring a known collection of AESI. Decision-theoretic methods are available (Lewis and Berry 1994; Stallard 1998), but require an elaborate mathematical framework and subjective determination of a loss function, making decisions at the end of the trial a difficult concept to quantify.

None of the methods discussed so far, frequentist or Bayesian, have been designed to be used with blinded data. To address this deficiency, we begin with a method like Thall and Simon (1994), but adapt it to be used with pooled, blinded data like Ball (Ball 2008; Ball et al. 2011; Yao et al. 2013; Wen et al. 2015). Using a Bayesian framework, Thall and Simon (1994) create statistical rules for single-arm clinical trials to compare a dichotomous efficacy endpoint of an experimental treatment to a standard therapy. The standard therapy is modeled on results from historical trials reflecting the uncertainty of the available information. The response rate of the experimental treatment is estimated by continuously updating a non-informative prior with data from the current trial. Ball et al. (2011) uses a simple Bayesian framework and a collaborative process to create continuous safety signals for a two-arm trial with a moderately informative prior on the combined event rate, which is updated with pooled, blinded data. We extend this idea with a fully Bayesian method that allows us to use a strong prior on the control rate and a separate weak prior on the treatment rate.

2 Safety Monitoring with Blinded Data

In 2002 Robert O'Neil, Director of Biostatistics at the Food and Drug Administration, declared that "statistical methodology has not been developed for safety monitoring to match that for efficacy monitoring" (O'Neil 2002). This problem has not gone away. We need quantitative methods that can be used to help guide our safety monitoring decisions.

Mills et al. (2006) carried out a thought-provoking review of HIV-AIDS trials. They found 82 full reports of trials, 10 (12 %) of which had been stopped early for harm. Though the medium sample size was 85, the maximum was over a thousand patients; and though the median study duration was a couple years, the maximum was over 5 years. A large number of patients were put at considerable risk, yet only six of these trials reported the use of a DMC and only one of them reported a plan for stopping early.

On the other hand, one trial was stopped early with a risk ratio for harm of 6.2. While this must have generated a lot of excitement, the decision to stop was based on a small number of unanticipated AEs; a total of 6 out of 38 patients (5 of 17 in the treatment arm and 1 of 21 in the control group). Statistical significance was not achieved, even without adjustment for multiple looks. The point estimate is unreliable, the confidence interval going from 0.8 to 40. This decision to stop was based on an ad hoc analysis with a lot of uncertainty. We propose a quantitative framework and a collaborative process that would improve conversations and help improve decisions about safety monitoring.

DMCs play a vital role in medical trials. To describe them briefly, there are three fundamental characteristics. First, they are independent from the ongoing collection and analysis of data. They are formed, by design, to provide an unbiased perspective. Second, they periodically evaluate accumulating unblinded data. They can use all of the available data, including treatment information, so that they can directly examine treatment effects and assess benefits to risks. And, third, they make recommendations about the continuing safe conduct of ongoing trials.

There is another key group of people responsible for monitoring the safety of clinical trial patients, which we will refer to as the Trial Leadership. This group includes anyone who has influence over the collection or analysis of trial data; such as, the Study Team and the Safety Management Team. Trial Leadership are invested in the ongoing collection and analysis of data and frequently evaluate accumulating blinded data. They do not have treatment information, as blinding prevents bias from their evaluations affecting the results of the study; however, they often have a better understanding of the rest of the data. Additionally, although the DMC make recommendations about stopping or continuing a trial, it is the Trial Leadership who take action to comply with investigational new drug (IND) reporting and make the final decision on whether or not to stop a trial. Clearly, Trial Leadership could benefit from objective statistical summaries that help them judge the strength of evidence contained in the blinded data.

Many methods have been developed that are referred to as stopping rules or decision rules. Not only is there a boundary that is defined with these rules, but the decision for what to do when that boundary has been crossed is also predetermined. This is problematic, as there are going to be more data that accumulate both inside and outside the study, before a decision is to be made. People in the design phase should not be establishing stopping rules and decisions to be implemented by a different group of people during the study who will have more knowledge. What we are proposing are just signals, mathematical summaries of the blinded data, about the incidence of specified AEs. They are not stopping rules, but signals to let the Trial Leadership know that they are getting to a place that made them uncomfortable when they designed the signals. It doesn't mean that they would have to stop. A trial would not be stopped unless the Trial Leadership made a fully informed decision to stop.

3 Method and Model

Safety monitoring is a dynamic process which can benefit from the flexibility of a Bayesian method designed for learning and decision-making. We present a unified framework for continuous safety monitoring which incorporates prior knowledge from earlier trials and assessment of the procedure's operating characteristics in order to gain a fuller understanding of the posterior distribution of AESI rates and to facilitate inferences with simple and easily interpretable probability statements.

Critical for designing useful safety signals is dynamic collaboration among subject matter experts from clinical, safety, and statistics groups. This multidisciplinary team must work together in discussing prior information and translating their collective knowledge into model parameters for safety signals that produce good operating characteristics. The essential elements of this process are described below:

1. Identify adverse events of special interest (AESI).
2. Determine plausible ranges for AESI rates in the control and treatment arms and establish the maximum acceptable rate for the treatment arm.
3. Translate AESI rate information and constraints into model parameters and carry out simulations for calibrating the operating characteristics of the safety signal.
4. Plot safety signal boundaries according to the safety monitoring plan.

The Safety Management Team would meet to identify a small collection of AESI for a focused search. AESI commonly arise due to drug class, Phase 1 observations, or other considerations. While statisticians would facilitate this part of the process, the safety group and clinical physicians would be the initial drivers due to their expert medical knowledge of internal and external studies in relevant patient populations. The team would also establish ranges of plausible AESI rates for the control and treatment arms, as well as a critical AESI rate for the treatment arm that would be concerning to Trial Leadership.

The statisticians would drive the remaining steps by translating the rate ranges and constraints identified in the previous steps into model parameters and priors, running a battery of simulated trials, and sharing the results with the rest of the team. The aim is to develop an active dialogue for calibrating the procedure to have good operating characteristics under a broad range of plausible circumstances. Once the team is comfortable with the results, but ideally before any patients have been enrolled, statisticians would determine and plot the maximum acceptable number of AESI by total number of patients enrolled for the whole trial. During the trial, the team could carry out exploratory analyses to more fully characterize the emerging safety profile from the full posterior.

4 Data Model

The known quantities at an interim analysis of blinded safety data are the number of enrolled patients, N, the number of patients observed with AESI, Z, and the expected proportion r of patients randomized to the treatment arm. The quantities of interest are the rates of AESI in the control arm θ_0 and the treatment arm θ_1. In particular, investigators may wish to know the posterior distribution of θ_1, the ratio θ_1/θ_0, or the absolute difference $\theta_1 - \theta_0$.

Note that while the randomization ratio r is known, the true number of patients assigned to each arm is unknown. Under simple randomization, the number of patients in the control arm M_0 has a Binomial(N, $1-r$) distribution, which leaves $M_1 = N - M_0$ patients in the treatment arm.

We assume that the AESI will occur soon after the treatment is administered, and we are not interested in counting the AESI after the first for each patient. In this case it is appropriate to model the number of patients with AESI in the control arm as $Y_0 \sim$ Binomial(M_0, θ_0) and the number of patients with AESI in the treatment arm as $Y_1 \sim$ Binomial(M_1, θ_1). If one or both of these assumptions are violated, another model such as a Poisson model may be more appropriate, as it can be used to model exposure time.

Specifying Beta priors for the AESI rates in each arm, the full model is then

Combined Arms	
$Z = Y_0 + Y_1$	
Control Arm	**Treatment Arm**
$Y_0 \mid M_0, \theta_0 \sim \text{Bionomial}\,(M_0, \theta_0)$	$Y_1 \mid M_1, \theta_1 \sim \text{Bionomial}\,(M_1, \theta_1)$
$M_0 \sim \text{Bionomial}\,(N, 1-r)$	$M_1 = N - M_0$
$\theta_0 \sim \text{Beta}\,(a, b)$	$\theta_1 \sim \text{Beta}\,(c, d)$

where a, b, c, and d are elicited hyperparameters.

A safety signal could be considered if the posterior probability, that θ_1 (the rate of AESI in the treatment arm) exceeds the maximum acceptable rate, had crossed some threshold. Similarly, the relative rate or rate difference could be used to trigger a safety signal.

5 Prior Specification and Operating Characteristics

In order to have a useful safety signal we must take full advantage of prior knowledge of AESI rates, especially in the control arm. Because we can only observe the combined AESI rate, the more we know about the AESI rate in the control arm, the more we can infer about the AESI rate in the treatment arm.

As the parameters of a Beta prior or posterior are easily interpretable as the number of prior events and non-events, specifying the hyperparameters directly is most feasible when the priors for the current trial are posteriors from a previous trial. For example, the prior for the control AESI rate may be a posterior from the Phase 3 trial of the control [Beta($a + 1$, $b + 1$), where a is the number of patients with AESI and b is the number of patients without AESI], while the prior for the treatment AESI rate could be the posterior from a previous trial of the treatment's development program (another Beta distribution, presumably less informative).

The operating characteristics of a method, or how it performs in a variety of possible settings, are essential to consider when choosing a method and its inputs. These operating characteristics may be determined through simulation prior to the start of the trial. The principal operating characteristic for safety monitoring is how frequently a procedure produces a safety signal under various values of the true AESI rates. The frequency with which a procedure produces a safety signal when the treatment is safe is the false positive error rate, and when the treatment is not safe this frequency is the power of the procedure. The probability threshold for θ_1 exceeding its maximum acceptable value is the most appropriate tuning parameter for influencing operating characteristics, though sensitivity of the operating characteristics to the prior specification should also be carefully considered.

6 Trial Procedure

Once the prior parameters, critical rate, and probability threshold have been set, then the maximum acceptable number of AESI for each number of enrolled patients can be computed. These maxima as well as the expected number of AESI (the average of the prior means weighted by the randomization ratio) can be plotted against the number of enrolled patients, as in Fig. 1. At each interim analysis a safety signal will be raised if the observed number of AESI exceeds the maximum acceptable number.

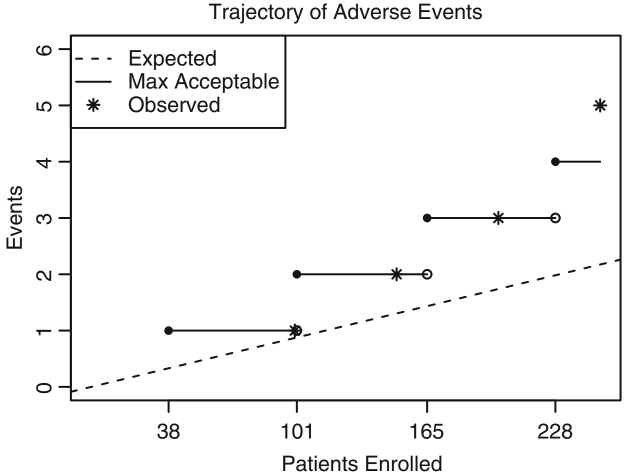


Fig. 1 The maximum acceptable number of adverse events versus the number of patients enrolled may be plotted before the trial begins, and observed counts marked at each interim analysis for straightforward inference

One thing to keep in mind is that these are only signals. We wouldn't necessarily decide to stop just because we crossed a probability threshold boundary. We want to be somewhat conservative in the sense that having a signal when we didn't need one is less of a concern than not having a signal when we should have had one. Having a signal when we didn't need one might use some additional resources but wouldn't be putting patients at a greater risk and wouldn't be jeopardizing the successful completion of the study.

These safety signals are not designed to replace DMCs, but rather to complement them. A DMC typically meets only a few times a year, so months could pass with no active review. Immediately after an assessment by the DMC we could be reasonably certain that the patients were safe, but as time goes on there would be a greater chance that a problem could have arisen. Blinded safety signals could help fill in this gap. They could also help prepare decision-makers for conversations with the DMC. Perhaps more importantly, blinded safety signals could be used as a mechanism to trigger an ad hoc meeting of a DMC, or the creation of a DMC in trials without one already established.

7 Summary

While DMCs should have unblinded access to all of the available information, Trial Leadership must conscientiously maintain their blind, so that together these two groups can fully protect patients from unsafe treatments without compromising trial integrity or otherwise interfering with the planned analyses.

We have developed a fully Bayesian method with a collaborative process that allows for continuous safety monitoring with blinded data and is ideal for learning and making decisions. At any time during the study, we could make easily interpretable posterior probability statements about arm-specific rates of AESI using the observed pooled numbers of events and patients. This method is not meant to satisfy every safety monitoring need, but rather to help Trial Leadership evaluate a small collection of AESI without unblinding the Study Team and without jeopardizing the successful completion of the study.

Ronald Fisher intended for p-values to be used as a part of the evidence, combined with other information, in the process of drawing conclusions from observations (Goodman 1999). Similarly, Neyman and Pearson (1928) developed their confidence intervals to be used as numerical measures of the data to help guide decision-makers. In the spirit of Fisher and Neyman and Pearson, we provide our blinded safety signals to be used as measures of the evidence in the blinded data that Trial Leadership can use, in combination with open information provided by the DMC, to evaluate the strength of evidence in all available data in order to make fully informed decisions that protect patients from unnecessary harm while allowing the trials to lead to conclusive results.

Acknowledgments We would like to thank Ann Eldred, Seemi Khan, Holly Read, Syed Islam, Karolyn Kracht, Shihua Wen and Jyotirmoy Dey of AbbVie for evaluating the process with an earlier version of our method.

References

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Guideline for Good Clinical Practice E6(R1). *ICH Harmonised Tripartite Guideline*. 1996.

European Commission. Detailed Guidance on the Collection, Verification, and Presentation of Adverse Reaction Reports Arising from Clinical Trials on Medicinal Products for Human Use. 2006.

US Department of Health and Human Services Food and Drug Administration. Safety Reporting Requirements for INDs and BA/BE Studies. *Guidance for Industry and Investigators*. 2012.

Crowe BJ et al. 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* 2009; **6:** 430–440.

Xia HA et al. Planning and core analyses for periodic aggregate safety data reviews. *Clinical Trials* 2011; **8:** 175–182.

Chuang-Stein C and Xia HA. The practice of pre-marketing safety assessment in drug development. *Journal of Biopharmaceutical Statistics* 2013; **23:** 3–25.

World Medical Association. Ethical Principles for Medical Research Involving Human Subjects. *World Medical Association Declaration of Helsinki*. 2006.

Council for International Organizations of Medical Sciences. International Ethical Guidelines for Biomedical Research Involving Human Subjects. 2002.

US Department of Health and Human Services Food and Drug Administration. Premarketing Risk Assessment. *Guidance for Industry.* 2005.

US Department of Health and Human Services Food and Drug Administration. Investigational New Drug Safety Reporting Requirements for Human Drug and Biological Products and Safety Reporting Requirements for Bioavailability and Bioequivalence Studies in Humans. *CFR*. 2010.

Huang L et al. A review of statistical methods for safety surveillance. *Therapeutic Innovation & Regulatory Science* 2014; **48 (1):** 98–108.

Wald A. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* 1945; **16 (2):** 117–186.

Goldman AI. Issues in designing sequential stopping rules for monitoring side effects in clinical trials. *Controlled Clinical Trials* 1987; **8:** 327–337.

Goldman AI and Hannan PJ. Optimal continuous sequential boundaries for monitoring toxicity in clinical trials: a restricted search algorithm. *Statistics in Medicine* 2001; **20:** 1575–1589.

Pocock SJ. Group sequential methods in the clinical design and analysis of clinical trials. *Biometrika*. 1977; **64 (2):** 191–199.

Pocock SJ. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics* 1982; **38:** 153–162.

O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35 (3):** 549–556.

Lan KKG and DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70 (3):** 659–663.

Anscombe FJ. Sequential Medical Trials. *Journal of the American Statistical Association* 1963; **58 (302):** 365–383.

Cornfield J. Sequential trials, sequential analysis and the likelihood principle. *The American Statistician* 1966; **20 (2):** 18–23.

Cornfield J. Recent methodological contributions to clinical trials. *American Journal of Epidemiology* 1976; **104 (4):** 408–421.

Herson J. Predictive probability early termination plans for phase II clinical trials. *Biometrics* 1979; **35:** 775–783.

Freedman LS and Spiegelhalter DJ. Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clinical Trials* 1989; **10:** 357–367.

Freedman LS, Spiegelhalter DJ and Parmar MKB. The what, why and how of Bayesian clinical trials monitoring. *Statistics in Medicine* 1994; **13:** 1371–1383.

Thall PF and Simon R. Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* 1994; **50 (2):** 337–349.

Thall PF Simon RM and Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine* 1995; **14:** 357–379.

Heitjan DF. Bayesian interim analyses of phase II cancer clinical trials. *Statistics in Medicine* 1997; **16:** 1791–1802.

Berry SM and Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics* 2004; **60:** 418–426.

Xia HA, Ma H and Carlin BP. Bayesian hierarchical modeling for detecting signals in clinical trials. *Journal of Biopharmaceutical Statistics* 2011; **21:** 1006–1029.

Lewis RJ and Berry DA. Group sequential clinical trials: a classical evaluation of decision-theoretic designs. *Journal of the American Statistical Association* 1994; **89 (428):** 1528–1534.

Stallard N. Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics* 1998; **54 (1):** 279–294.

Ball G. Continuous safety screens for randomized controlled clinical trials with blinded treatment information. *(Doctoral Dissertation)* 2008.

Ball G, Piller LB and Silverman MH. Continuous safety monitoring for randomized controlled clinical trials with blinded treatment information. *Contemporary Clinical Trials* 2011; **32:** S1–S10.

Yao B et al. Safety monitoring in clinical trials. *Pharmaceutics* 2013; **5:** 94–106.

Wen S, Ball G, Dey J. Bayesian monitoring of safety signals in blinded clinical trial data. *Annals of Public Health and Research* 2015;2(2):1019–1022.

O'Neal RT. Regulatory perspectives on data monitoring. *Statistics in Medicine* 2002; **21:** 2831–2842.

Mills E, Cooper C, Wu P, Rachlis B, Singh S, Guyatt GH. Randomized trials stopped early for harm in HIV/AIDS: a systematic survey. *HIV Clinical Trials* 2006; **7:** 24–33.

Goodman SN. Toward evidence-based medical statistics: the p-value fallacy. *Annals of Internal Medicine* 1999; **130:** 995–1004.

Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika* 1928; **20:** 175–240.

Part VI
Statistical Applications in Nonclinical and Preclinical Drug Development

Design and Statistical Analysis of Multidrug Combinations in Preclinical Studies and Phase I Clinical Trials

Ming T. Tan, Hong-Bin Fang, Hengzhen Huang, and Yang Yang

Abstract Multidrug combination is an important therapeutic approach for cancer, viral or microbial infections, hypertension and other diseases involving complex biological networks. Synergistic drug combinations, which are more effective than predicted from summing effects of individual drugs, often achieve increased therapeutic index. Because drug-effect is dose-dependent, multiple doses of an individual drug need to be examined, yielding rapidly increasing number of combinations and a challenging high dimensional statistical modeling problem. The lack of proper design and analysis methods for multi-drug combination studies have resulted in many missed therapeutic opportunities. Although systems biology holds the promise to unveil complex interactions within biological systems, the knowledge on network remains predominantly topological until very recently. This article summarizes recent work on efficient maximal power experimental designs on multidrug combinations, and statistical modeling of the resulting data. The design and analysis of vorinostat and cytarabine combination study is presented to illustrate the approach. We then introduce a model based adaptive Bayesian phase I trial design for drug combinations utilizing the modeling concept. To tackle the challenging problem of combinations of more than three drugs, we present a novel two-stage procedure starting with an initial selection by utilizing an in silico model built upon experimental data of single drugs and current systems biology information to obtain maximum likelihood estimate.

Keywords Clinical trial design • Biological networks • Drug combinations • Experimental design • Maximal Power design • Statistical modeling • Synergy analysis

M.T. Tan (✉) • H.-B. Fang • H. Huang
Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University
Medical Center, 4000 Reservoir Rd NW, Washington, DC, USA
e-mail: mtt34@georgetown.edu

Y. Yang
Division of Biometrics 1, Office of Biostatistics, Center for Drug Evaluation and Research,
US Food and Drug Administration, Silver Spring, MD, USA

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_18

1 Introduction

In the past decade the identification of a variety of novel signal transduction targets amenable to therapeutic intervention has revolutionized the approach to cancer therapy. These targets were identified based on improved understanding of the molecular mechanisms of action of second messengers, other components of signal transduction pathways and systems biology. These advances have also made available large number of potential agents and call for new quantitative approaches for combination therapy (Xavier and Sander 2010; Fitzgerald et al. 2006; Hopkins 2008), which then motivated the development of design and analysis methods for three drugs (Tan et al. 2009; Fang et al. 2015).

Multi-drug combination is an important therapeutic approach for diseases such as cancer, viral or microbial infections, hypertension and other diseases involving complex biological pathways. Synergistic drug combinations, which are more effective than expected from summing effects of individual drugs, offer the potential for improved therapeutic index. Because drug-effect is dose-dependent, multiple doses of an individual drug often need to be examined, yielding rapidly increasing number of combinations that prohibit experimentation, and yielding a challenging high dimensional statistical problem. The lack of proper design and analysis methods for multi-drug combination studies have resulted in suboptimal utilization research resource and missed therapeutic opportunities.

The past decade has seen significant progresses in developing proper design and analysis methods for multi-drug combination studies have increased the chances of identifying optimized combinations for further therapeutic opportunities for combinations of two, three, or more drugs utilizing optimized designs and systems biology (see, e.g., Tan et al. 2003; Fitzgerald et al. 2006; Fang et al. 2008, 2015, 2016; Calzolari et al. 2008) as well as adaptive phase I clinical trial designs that attempt to identify the best possible maximum tolerated doses through modeling of the joint dose-toxicity relationship (see, e.g., Yuan and Yin 2008; Yin and Yuan 2009a, b; Yang et al. 2016). The non-model based designs include the approach using a partial order of toxicity discussed in Wages et al. (2011) and a two-dimensional extension of the biased coin design (Sun and Braun 2015). These approaches are a welcome step-forward as they all have done away with the problematic assumption that the dose-limiting toxicity (DLT) increases monotonically with increasing doses. It is known this assumption is reasonable in single agent phase I trials, it may not hold in drug combinations since the ordering of the probabilities of DLT of these combinations is not known at the design stage of the trial.

This article is to review the development of the vorinostat (SAHA) combinations for leukemia from nonclinical studies to clinical trials. We then summarize recent statistical methods motivated by and used in the vorinostat development as well as lessons learned moving the therapy to clinic. Specifically, we present an efficient experimental design on selected multi-drug combinations, statistical modeling of the resulting data and the proof of its statistical properties. Drawing experience from the vorinostat studies, we present an adaptive Bayesian trial design for multidrug combinations with interaction modeling and an optimized design for multidrug combinations. We also discuss applications and areas that are likely to assume an important role in future drug discovery and development research, such as ways of dealing with the difficult high dimensional problem with multidrug combinations utilizing in silico models that integrate statistical modeling, experimental data of single drugs and current systems biology approach.

The rest of the article is organized as follows. Section 1 introduces the combination study of vorinostat and the efficient experimental design, the maximal power method for drug combinations using Loewe additivity. Section 2 describes the analysis of vorinostat and cytarabine (ara-C) combinations, how it affects the clinical trial design, and the clinical trial results, how it has impacted the development of methodology to design multidrug combination studies and how a potentially useful approach for phase I trial design that utilizes the modeling approach can be derived. Section 3 introduces the Bayesian adaptive phase I trial deign for drug combinations while modeling the interaction based on Bliss independence. Section 4 introduces current work on multidrug combinations of three drugs. Section 5 presents a novel two-stage procedure starting with an initial selection by utilizing an in silico model built upon experimental data of single drugs and current systems biology information to obtain maximum likelihood estimate by integrating modern statistical methods and systems biology approaches. We conclude with a discussion on the future of this field.

2 Vorinostat Combination Studies and Maximal Power Design

Vorinostat (suberoylanilide hydroxamic acid, SAHA) is a small molecule histone deacetylase (HDAC) inhibitor that is currently the most potent HDAC inhibitor available clinically. The vorinostat combination trial is a phase I trial to determine the maximum tolerated dose of vorinostat used in combinations of the mainstay of anti-leukemia chemotherapeutic agents (Gojo et al. 2013). To investigate the potential activity of the combination, extensive preclinical in vitro cytotoxicity studies on vorinostat combined with ara-C and etoposide as well as the sequence of administration have been performed to test the interaction (synergy or antagonism) of the combination for treating acute leukemia (Shiozawa et al. 2009). Ara-C is one of the most active agents available for treating acute leukemia. Etoposide has been shown to be an effective anti-leukemia agent, particularly when given in combination with other chemotherapeutic agents. It exerts its effects by interfering with topoisomerase II activity, binding to and stabilizing the covalent linkage between topoisomerase II and DNA, and inhibiting the re-ligation of the resultant DNA double strand breaks.

Experimental approaches to characterizing combination therapy typically involve determining dose–response curves for inhibitors individually and in combination. When experimental dose–response data match the predictions of Loewe additivity, the inhibitors are said to be additive (corresponding to the zero-interaction case); greater than predicted potency indicates synergism (positive interaction); and lower potency argues for antagonism (negative interaction). With different dose–response curves of individual inhibitors, various measurements for combinations have been developed according to Loewe additivity. Based on the median-drug effect analysis method which assumes that two drugs alone or in combination will result in sigmoid concentration-effect curves, Chou and Talalay (1984) defined a combination index. Assuming the dose–response curves of individual drugs can be characterized by Hill models, Greco and his colleagues proposed an equation to characterize interactions of two drugs (Greco et al. 1995). Peterson and Novick (2007) derived a nonlinear blending measurement for the assessment of combination drug synergy.

The very first set of experiments were conducted based on one fixed dose ratio of vorinostat of ara-C and etoposide, which missed the complexity of the interaction and precluded attainment of data on important interactions among these agents. Consequently, we have designed the study to include various combinations that are determined based on an efficient statistical design so that the statistical power to demonstrate the departure from additivity is maximized (Tan et al. 2003, 2009; Fang et al. 2008). Indeed it was shown that vorinostat interacted additively or synergistically with etoposide; but not with ara-C because vorinostat diminished cells in cell cycle S-phase, where cells are most vulnerable to ara-C toxicity (Shiozawa et al. 2009). However, the sequential administration of vorinostat followed by ara-C with a 72-h interval demonstrated synergy, where the time between administration of vorinostat and ara-C allows cells to re-enter into S-phase (Shiozawa et al. 2009; Gojo et al. 2013). This article focus on the vorinostat plus ara-C combination study to demonstrate the methodology.

To introduce the *maximal power design* (MPD) to detect departures of additivity, i.e., detecting synergy or antagonism, we first review how additivity, the expected dose effect when the two drugs are considered additive, is defined. There are two commonly used definitions, the Bliss independence and the Loewe additivity (Berenbaum 1989; Fitzgerald et al. 2006), although validity of this model as a universal reference model has been questioned (Greco et al. 1995). The Loewe additivity assumes that two inhibitors exert their effect through a similar mechanism (e.g., pathway), where the effects of each inhibitor and the combination are related through equipotent dose ratios. Bliss independence, however, assumes that the two inhibitors act through independent mechanisms (e.g., multiple pathways), in which combination therapy is represented as the union of two probabilistically independent events. The Loewe additivity correctly predicts the trivial case in which the two drugs are actually the same compound, i.e., drug A and a dilution of it are additive (Berenbaum 1989).

The Loewe additivity is embodied in the isobologram method for characterizing departures from additivity. To describe the joint action of two drugs A and B at a specific dose level, the additivity of Loewe (1955) is based on single drug dose-effect and is defined by the following isobole equation

$$\frac{x_A}{X_A} + \frac{x_B}{X_B} = \tau \tag{1}$$

where x_A and x_B are doses of the constituent drugs A and B of the combination needed to yield a given level of effect, e.g., 50 % inhibition (ED_{50}), or 50 % death in experiment animals (LD_{50}), where X_A and X_B are the doses needed for each drug alone to yield the level of effect. The τ is called the interaction index of the drugs A and B at the combination (x_A, x_B).When $\tau = 1$, the drugs A and B is additive (zero-interaction) at the combination (x_A, x_B); when $\tau < 1$, they are synergistic, namely, the combination (x_A, x_B) is more effective than expected from their single drug dose–response curves, otherwise ($\tau > 1$), they are antagonistic.

Let the dose–response relationships for individual drugs A and B be $y = f_A\,(X_A)$ and $y = f_B\,(X_B)$ respectively. Denote the combination dose-effect (response) by $f_{com}(x_A, x_B)$, and with (1) we have

$$\begin{aligned} f_{com}\,(x_A, x_B) &= f_A\,(X_A) = f_A\,(\tau X_A) + [f_A\,(X_A) - f_A\,(\tau X_A)] \\ &= f_A\left(x_A + \tfrac{X_A}{X_B} x_B\right) + [f_A\,(X_A) - f_A\,(\tau X_A)]\,. \end{aligned} \tag{2}$$

The term $[f_A\,(X_A) - f_A\,(\tau X_A)] = 0$ if the drugs are additive ($\tau = 1$). Then, the regression line for the combination with additive action of two drugs is $y = f_A\,(x_A + \rho_B\,(X_B)\,x_B)$ where the relative potency $\rho(X_B)$ is a function of X_A and X_B, $\rho\,(X_B) = f_A^{-1} f_B\,(X_B)\,/X_B$. As show in Fang et al. (2008), the potency $\rho(X_B)$ is generally not a constant, the additive model (2) has no closed form.

Since we typically do not know much about the joint effect of the combinations before experiments, we have proposed a general semiparametric model for the joint effect of the constituent drugs (Tan et al. 2003),

$$y = f_A\,(x_A + \rho\,(X_B)\,x_B) + f\,(x_A, x_B) + \varepsilon \tag{3}$$

where $f(x_A, x_B)$ is an unspecified function since the term $[f_A\,(X_A) - f_A\,(\tau X_A)]$ in (2) is a function of (x_A, x_B), ε is the error term due to variation in experiments and is assumed to be normally distributed with mean 0 and variance σ^2. Then, testing the additive action of the two drugs is equivalent to testing the null hypothesis $H_0 : f = 0$. Suppose that there is a one-to-one invertible transformation from $(x_A, x_B) \to (z_1, z_2)$ such that $f_A\,(x_A + \rho\,(X_B)\,x_B) = (\text{or} \approx)\, g_1\,(z_1) + g_2\,(z_2)$, where the functions g_1 and g_2 are linearly independent, an F-test is derived using a lack of fit type of test with the sum of squares with and without the term f. Specifically, let the m mixtures $z^{(1)}, \ldots, z^{(m)}$ be in the experimental domain. Assume that there are n_i experiments at the dose-level $z^{(i)} = \left(z_1^{(i)}, z_2^{(i)}\right)^T$ with corresponding responses $y_{ij}\,(j = 1, \ldots, n_i; i = 1, \ldots, m)$ .. Denote $n = n_1 + \cdots + n_m$, $\boldsymbol{y}$ the $n \times 1$ vector with elements y_{ij} ordered lexicographically, Z the $m \times 2$ matrix with i-th row $(g_1(z_1^{(i)}), g_2(z_2^{(i)}))$. Let $V = UZ\big(Z^T U^T UZ\big)^{-1} Z^T U^T$, $J = U\big(U^T U\big)^{-1} U^T$, and the $n \times m$ matrix $U = diag\,(\mathbf{1}_{n_1}, \cdots, \mathbf{1}_{n_m})$. Then, if the hypothesis H_0 is true, the statistic

$$F = \frac{\mathbf{y}^T (J - V)\mathbf{y}/(m-2)}{\mathbf{y}^T (I - J)\mathbf{y}/(n-m)}, \tag{4}$$

has a central F-distribution with degrees of freedom m–2 and n–m (Tan et al. 2003; Fang et al. 2008).

The question is which combinations should be chosen for experiment to demonstrate synergy, antagonism or additivity efficiently and with adequate statistical power. The MPD then utilizes individual dose response data and uniform measures (Fang and Wang 1994) to select a moderate number of combinations with a preset number of replications for experimentation (Tan et al. 2003, 2009; Fang et al. 2008). This method maximizes the minimum (among potential forms of departures from additivity) power of the F-test in (4) to detect departures from the additive action of drugs.

Although there exists conceptual statistical work, e.g., the maximal power F-test, for finding doses and sample sizes needed to detect departures from additivity. However, the method depends on dose–response shapes of individual drugs, namely, different classes of drugs of different dose–response shapes require different derivations for sample size and dose finding.

Upon completion of the experiments, the F-statistic (4) is to test the hypothesis of the additive action of two drugs and calculate the p-value of the F-test. If the p-value is greater than 0.05, we can accept the hypothesis of the additive action of two drugs. Otherwise, we calculate the interaction index (τ) as follows. Let y_{ij} be the j-th response at $(x_A^{(i)}, x_B^{(i)})$. With the single dose–response curves, the interaction indexes at $(x_A^{(i)}, x_B^{(i)})$ are

$$\tau_{ij} = \frac{x_A^{(i)}}{f_A^{-1}\left(y_{ij}\right)} + \frac{x_B^{(i)}}{f_B^{-1}\left(y_{ij}\right)}, \quad j = 1, \ldots, k; \;\; i = 1, \ldots, m. \tag{5}$$

The method of two-dimensional B-splines (thin plate splines) is employed to estimate the interaction index surface $\tau = h\left(x_A, x_B\right)$ (Fang et al. 2008).

In the vorinostat plus ara-C combination study, we first considered the experimental design. Based on the single experiments of inhibiting HL-60 cell line, the estimated dose–response curves of vorinostat and ara-C are

$$\begin{aligned} y &= 51.04 - 20.88 \log\left(X_A - 0.05\right), && X_A \in [0.1\ \mu\text{M}, 10\ \mu\text{M}], \\ y &= 9.22 - 10.17 \log\left(X_B\right), && X_B \in [0.003\ \mu\text{M}, 0.6\ \mu\text{M}], \end{aligned} \tag{6}$$

respectively, where y is the viability (% of control). The corresponding ED_{50} of vorinostat and ara-C are 1.101 μM and 0.021 μM, respectively. To investigate the synergy of vorinostat and ara-C against HL-60, we used the MPD for the mixture experiments. The variance is estimated to be 1006.416 based on the pooled observations from the single drug experiments. The MPD design yields 18 combinations with five replicates at each combination (Table 1). The design has 80 % statistical power to detect at least a 15 % difference in viability between the

Table 1 Mixtures of vorinostat and ara-C

Vorinostat (μM)	ara-C (μM)	Vorinostat (μM)	ara-C (μM)	Vorinostat (μM)	ara-C (μM)
0.137	0.162	2.576	0.357	2.483	0.021
0.568	0.586	1.186	0.045	5.005	0.048
0.321	0.050	2.804	0.167	1.934	0.006
1.033	0.295	0.875	0.011	4.737	0.012
0.247	0.008	2.772	0.067	1.127	0.003
1.239	0.129	0.305	0.003	4.210	0.003

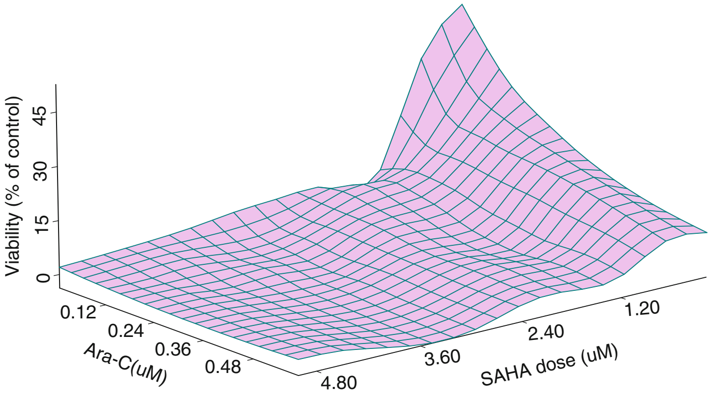


Fig. 1 Response surface of the sequential combination of vorinostat (SAHA) with ara-C

predicted additive values and the observed values at a significance level of 5 %. Then, cells are exposed to these select combinations and the cytotoxicity of this combination is determined.

In the sequentially combination experiments of vorinostat with ara-C against HL-60, the dose ranges are from 0.137 to 5.005 μM for vorinostat and from 0.003 to 0.586 μM for ara-C. Of total 108 observations, the maximum viability is 78.87 % and the minimum viability is 0.027 %. The mean is 15.00 % and the standard error is 17.373. Figure 1 shows the response surface of the combination of vorinostat with ara-C against HL-60. The F-test (4) shows that we reject the null that vorinostat with ara-C against HL-60 has additive action ($F_{16,90} = 16.85$, p-value < 0.0001). To explore the interaction of vorinostat with ara-C, we estimated the interaction index surface $\tau = h(x_A, x_B)$ using thin plate splines (Fang et al. 2008). Figure 2 shows the contour plot of combination index surface such that when the dose of vorinostat is less than 0.4 μM or both the doses of vorinostat and ara-C are higher, the joint action is additive. The maximum synergy actions occur at the dose of vorinostat between 1.2 and 2.5 μM and the dose of ara-C between 0.003 and 0.3 μM.

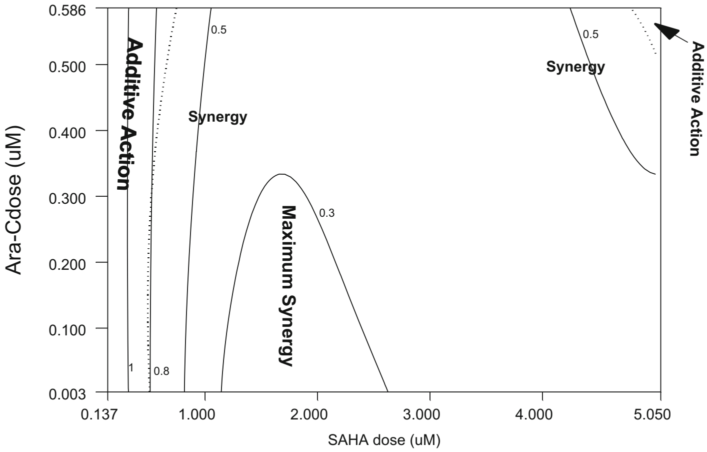


Fig. 2 Contour plot of combination index surface of vorinostat (SAHA) and ara-C sequential combination. The *dotted lines* indicate the 95 % confidence surfaces for additive action (the combination index = 1)

Based on the preclinical results, a phase I trial was planned. In principle, modeling the toxicity interaction appropriately would add to the efficiency and result in a better trial design. However, at the time of designing the phase I protocol, none of the methods were ready for clinical trial protocol developments. We designed the phase I trial escalating the dose of vorinostat while having the fixed doses of ara-C (1–2 g/m^2 due to patient age) and etoposide (100 mg/m^2) on days 11–14. Twenty-one patients with acute myelogenous leukemia (AML) were enrolled in the trial, and the maximum-tolerated dose (MTD) was established to be vorinostat 200 mg twice a day orally. Of 13 patients with high-risk leukemia treated at the maximum-tolerated dose of vorinostat (200 mg, orally, twice a day), six obtained a complete remission (CR) with median duration of 7 months. The relatively high CR rate in this poor-risk acute myelogenous leukemia patient group warrants further study (Gojo et al. 2013).

However, there are two missed opportunities for this study: a suboptimal clinical trial design had been used where only the dose of vorinostat was escalated, and a suboptimal study design with fixed dose of one drug for the three drug combination had been used. In the following two sections, we present both a Bayesian adaptive phase I trial design that would have been useful in identifying the maximum tolerated doses in three dimensions; and the maximal power design for combinations of three drugs that would have been utilized in the Vorinostat study had these methods been available then.

3 Bayesian Adaptive Phase I Trial Design for Drug Combinations

Different from single agent trials, the interaction effect between two drugs may have a significant impact on the joint toxicity probability of the dose combination. Independent, synergistic or antagonistic effects are the different states of interactions. The independence model implies that the two drugs have no apparent interaction with the respect to the toxicity. The *Bliss independence criterion* has been used in describing the joint action in toxicity for two agents (Goldoni and Johansson 2007). Its main assumption is that two or more drugs act independently from one another (Greco et al. 1995; Bliss 1939; Berenbaum 1989). Let $P(A)$ and $P(B)$ be the marginal toxicity probabilities of drugs A and B, respectively. If A and B are independent, the probability of no toxicity in the combination of A and B is

$$1 - P\left(A \cup B\right) = \{1 - P(A)\}\{1 - P(B)\} \tag{7}$$

Thus, the joint probability of toxicity $g(x_A, x_B)$ at combination (x_A, x_B) has the form

$$g\left(x_A, x_B\right) = 1 - \{1 - g\left(x_A, 0\right)\}\{1 - g\left(0, x_B\right)\}, \tag{8}$$

where $g(x_A, 0)$ and $g(0, x_B)$ are the marginal toxicity probabilities of drug A and drug B, respectively. When $g\left(x_A, x_B\right) > 1 - \{1 - g\left(x_A, 0\right)\}\{1 - g\left(0, x_B\right)\}$, drug A and drug B at combination (x_A, x_B) have Bliss synergy of toxicity. For Bliss antagonism, the inequality is reversed. The Bliss antagonism results in lowering toxicity at a given drug combination. To specify the toxicity response $g(x_A, x_B)$, we proposed a factorial type Bliss model that allows mixed interaction profile for the combination therapy using the drugs A and B on the binary toxicity outcome. The probability of toxicity is modeled as follows:

$$g\left(x_A, x_B, \theta\right) = 1 - \exp\left(-\alpha x_A - \beta x_B\right)^{f(\gamma_1, \gamma_2, x_A, x_B)}, \tag{9}$$

where $\alpha > 0$, $\beta > 0$ and γ_1, γ_2 are parameters to be estimated. The function $f(\gamma_1, \gamma_2, x_A, x_B)$ is used to measure the degree of synergy or antagonism of the different dose combinations. We proposed the following form

$$f\left(\gamma_1, \gamma_2, x_A, x_B\right) = \exp\left(x_A x_B\left(\gamma_1 x_A + \gamma_2 x_B\right)\right). \tag{10}$$

The model satisfies the conditions that if $x_B = 0$, then $g\left(x_A, 0\right) = 1 - \exp\left(-\alpha x_A\right)$, which is the toxicity model of single drug A. Similarly, when $x_A = 0$, then $g\left(0, x_B\right) = 1 - \exp\left(-\beta x_B\right)$, the toxicity model reduces to that of the single drug B. Then, the single drug case becomes the convention exponential toxicity model. It captures antagonism when $f\left(\gamma_1, \gamma_2, x_A, x_B\right) < 1$, independence when $f\left(\gamma_1, \gamma_2, x_A, x_B\right) = 1$ and synergy when $f\left(\gamma_1, \gamma_2, x_A, x_B\right) > 1$. Thus, we call $f(\gamma_1, \gamma_2, x_A, x_B)$ the interaction function.

Based on the toxicity model (9), we proposed a novel method to find the *maximum tolerated region* (MTR) consisting of the doses that have the posterior mean toxicity probabilities below the target toxicity probability φ,

$$\mathrm{MTR} = \{(x_A, x_B) : \pi\left((x_A, x_B)\,;\theta\right) \leq \varphi\} \tag{11}$$

with minimum patient number for a given target probability. The method recognizes that there may exist multiple MTDs with drug combinations and addresses the issue directly. It integrates the toxicity interaction of two drugs and Bayesian adaptive dose-finding algorithm (Yang et al. 2016). The goal is to bring the trial to dose combinations where there may be antagonistic behavior among the drugs that the patients can be safely assigned to relatively high dose of individual drugs which otherwise would not be possible in single drug scenario. Thus, patients can be exposed to doses with high efficacy without experiencing significant toxicity. Let $x_A = \{a_1, \ldots, a_I\}$ and $x_B = \{b_1, \ldots, b_J\}$ be the specific dose levels of drugs A and B, respectively. For given dose levels of drug A at x_A and drug B at x_B, the rescaled interaction function can be defined as:

$$v\left(x_A, x_B\right) = \frac{f\left(\gamma_1, \gamma_2, x_A, x_B\right)}{f\left(\gamma_1, \gamma_2, x_A, x_B\right) + 1}. \tag{12}$$

The goal will be to locate a dose combination by minimizing a combination of the interaction function $v(x_A, x_B)$ and the toxicity probability $g(x_A, x_B)$ subject to the constraint that the toxicity probability is no more than a pre-specified value. We define the objective as a convex combination of the probability of toxicity at dose (x_A, x_B) and the interaction $v(x_A, x_B)$ at that dose,

$$U_\lambda\left(x_A, x_B\right) = \lambda g\left(x_A, x_B\right) + (1-\lambda)\, v\left(x_A, x_B\right), \tag{13}$$

where $0 < \lambda < 1$. The choice of λ reflects how much emphasis one would like to put on having more allocation at antagonistic combinations. Toxicity probability and interaction function are considered jointly through the objective function with the relative contribution of each component controlled by the weight λ. Smaller values of U would indicate smaller values of the standardized interaction leading to more antagonism and smaller toxicity probability. Therefore, our Bayesian adaptive dose-finding design is developed with the goal of minimizing the objective function. The objective function is evaluated based on the measurement of the mean squared error (MSE) of the toxicity probability estimate and the amount of interaction that patients really experienced. We conducted extensive empirical studies to evaluate possible λ values over several plausible scenarios. We recommend the choice of λ = 0.5 which suggests equal contribution of toxicity probability and the allowance of interaction. The next dose combination may be chosen to minimize the posterior expectation of the objective function given the current data Z_n,

$$\boldsymbol{x}_{n+1} = \left(a_i, b_j\right) = \arg\min E\left\{U_\lambda\left(\boldsymbol{x}, \theta\right) \middle| Z_n\right\}. \tag{14}$$

The dose finding algorithm and simulation studies of the design properties are given in Yang et al. (2016). The simulation studies under various scenarios demonstrate that the proposed design performs satisfactorily with expected operating characteristics. In particularly, the sample size in this proposed method is more favorable than that in existing methods based on the simulation results.

4 Maximal Power Design for Three-Drug Combinations

As we have shown in Sects. 2 and 3, the preclinical experiments of the vorinostat combinations have been done pairwise, vorinostat + etoposide at a fixed ara-C dose, and vorinostat + ara-C at a fixed etoposide dose. However, the optimal design was not developed early enough to be utilized in the development of vorinostat combinations. Indeed, to our knowledge, the literature in experimental design method for three drug combinations is sparse. Tan et al. (2009) derived the potential experimental design of combinations with varying doses in all three constituent drugs determined based on the MPD by extending the method summarized in Sect. 2, and proposed a sample size formula and a MPD to detect synergy in combination studies of three drugs when each of three drugs has a log-linear dose–response. Since the design depends on the shapes of the dose–response curves of the constituent agents, combination studies based on the linear and log-linear individual dose–response curves necessitate different mathematical joint effect model and generating uniformly scattered points in a tetragon area (Tian et al. 2009). That was the first time to our knowledge that a three drug combination experiment was designed through a search of the three drug dose region.

Note that critical to the uniform design method is to be able to derive the approximate decomposition of the additive model, and this becomes more difficult with three drugs. Tan et al. (2009) has derived the MPD for combinations of common cytotoxic agents whose individual dose response is log-linear or observe the Hill model. The log-linear dose–response curve represents a wide class of drugs including antimetabolites, antibiotics, interferons, growth factors, neuropeptide Y, phorbol esters, narcotics and neuronal agonists, hepatotoxins, and cromoglycate. The combination of vorinostat combined with ara-C and etoposide against HL-60 is also considered there.

To illustrate the methods of experimental design for three-drug combination studies, we consider the experiment to determine the effects of pre-administration of vorinostat on the pharmacokinetics of ara-C and etoposide against the leukemia cell line HL-60 (Shiozawa et al. 2009). In the experiments for single agents, we have 56 observations with doses ranging from 0.1 to 10 μM for vorinostat, 56 observations with doses ranging from 0.003 to 0.6 μM of ara-C, and 64 observations with doses ranging from 0.01 to 10 μM of etoposide. Then, the single dose–response curves for ara-C, etoposide and vorinostat are estimated to be

$$\begin{aligned} y\left(X_A\right) &= 4.80 - 12.76 \log\left(X_A\right), \\ y\left(X_B\right) &= 41.52 - 13.02 \log\left(X_B\right), \\ y\left(X_C\right) &= 54.55 - 23.98 \log\left(X_C\right), \end{aligned} \tag{15}$$

respectively, where y is the 100× viability, and X_A, X_B and X_C are the doses of ara-C, etoposide and vorinostat respectively. The potency of etoposide relative to ara-C is $\rho_0\left(X_B\right) = 0.0563 X_B^{0.0204}$ and the potency of vorinostat relative to ara-C is $\rho_1\left(X_C\right) = 0.0203 X_C^{0.8793}$, which show that these relative potencies are non-constant and depend on dose. The predicted additive model at (x_A, x_B, x_C) is

$$y\left(x_A, x_B, x_C\right) = 4.80 - 12.76 \log\left(x_A + 0.0551 \psi^{0.0427} x_B + 0.0203 \psi x_C\right), \tag{16}$$

where ψ is determined by

$$\psi = \left(49.3483 x_A + 2.7204 \psi^{0.0427} x_B + \psi x_C\right)^{0.4679}.$$

An approximate additive model (16) is given by

$$\begin{aligned} y_{\text{appr}}\left(x_A, x_B, x_C\right) = {} & 4.80 - 12.76 \log\left(z_1\right) - 12.76 \log\left[(1 - 0.0563)\, z_2 + 0.0563\right] \\ & -12.76 \log\left[(1 - 0.3601)\, z_3 + 0.3601\right], \end{aligned} \tag{17}$$

where

$$\begin{cases} z_1 = x_A + 0.9798 \psi^{0.0427}\left(x_A, x_B, x_C\right) x_B + \psi\left(x_A, x_B, x_C\right) x_C \\ z_2 = x_A\left[z_1 - 0.6399 \psi\left(x_A, x_B, x_C\right) x_C\right]^{-1} \\ z_3 = \psi\left(x_A, x_B, x_C\right) x_C / z_1 \end{cases} \tag{18}$$

To obtain MPD for testing the joint action of ara-C, etoposide and vorinostat, the dose range is chosen such that the endpoint, 100 × viability, is from 20 to 80 for ara-C. Then, the total dose ranges from 0.0028 to 0.3038 μM in ara-C. The pooled variance from the three single drug experiments is 988.422. For a meaningful difference η of 15 (100 × viability) and five replications for each mixture, with type I error rate 0.05 and power 0.80, we need study 21 mixtures in the experiment in order to detect synergy/antagonism in the combination of ara-C, etoposide and vorinostat (total 105 experiments). With the algorithm given in Tan et al. (2009), we get 21 points in domain $\left\{(z_1,\ z_2,\ z_3)^T :\ \ 0.0028 < z_1 < 0.3038,\ z_2 > 0,\ z_3 > 0,\ z_2 + z_3 < 1\right\}$. According to the inverse transformation of (18), 21 mixtures of these three drugs for experiments are given in Table 2, of which the doses of etoposide and vorinostat are $16.78149(x_B^{(i)})^{0.98}$ and $7.961724(x_C^{(i)})^{0.5321}$ respectively, because of the total dose range according to ara-C.

Furthermore, the method needs to be modified to allow one or more of the individual dose response curve being not log-linear. For example, in the combination

Table 2 Twenty-one mixtures of ara-C, etoposide and vorinostat for combination experiment

Exper. #	ara-C (μM)	Etoposide (μM)	Vorinostat (μM)	Exper. #	ara-C (μM)	Etoposide (μM)	Vorinostat (μM)
1	0.0012	0.0933	0.4749	12	0.1248	0.6983	0.3619
2	0.0059	1.2214	1.4168	13	0.0094	2.7681	0.9663
3	0.1883	0.4162	1.1650	14	0.0850	0.9432	1.7338
4	0.0005	0.2020	1.4923	15	0.0450	1.7054	1.8895
5	0.0283	0.6632	0.3619	16	0.0224	1.2908	2.6776
6	0.0204	0.2379	1.5646	17	0.0052	0.5223	0.4749
7	0.0523	0.5928	0.7225	18	0.0065	2.2552	2.5879
8	0.0138	0.0933	0.4749	19	0.0692	3.0748	0.9663
9	0.0388	0.5928	1.6679	20	0.0471	1.1172	3.0914
10	0.0475	1.1867	1.0697	21	0.0981	2.0494	2.0626
11	0.0081	0.3451	2.6776				

study of three anti-cancer agents PD184, HA14-1 and CEP3891 in myeloma H929 cell line. PD184 is a highly potent and selective noncompetitive MEK inhibitor. HA14-1 is a small, cell-permeable nonpeptidic ligand that binds to the Bcl-2 surface pocket and blocks its biological action. Similar studies involving three agents have been designed and analyzed in a pairwise fashion, namely, studying the combinations of any two of the three separately. This is clearly suboptimal, not only it is potentially more costly but also the analysis results are hard to interpret and may not reflect the real optimal dose of the three agents in combination (Pei et al. 2003, 2004). Fang et al. (2015) extended the MPD and derived the design and analysis in this case with mixed linear and log-linear dose response curves. The experiment of varying doses of all three drugs based on the MPD was implemented for the first time to our knowledge.

5 Multidrug Joint Response Modeling with Systems Biology

Increasing the number of agents in a combination may provide better outcomes. However, even with six drugs, each with only six doses, the number of potential combinations reaches 46,656. The exponential increase in number of combinations with the number of drugs makes laboratory testing difficult. Consequently, most work in multidrug combinations is conceptual. Calzolari et al. (2008) utilize a deterministic model and network information to develop a search algorithm. Furthermore, despite the biological advances mentioned above and the importance of multi-agent combinations, current methods are mostly topological as opposed to quantitative, and do not account for high dimensionality and proper model assumptions (Krzywinski and Altman 2014; Ashton 2015). Recently we proposed a novel two-stage procedure utilizing an initial selection by utilizing an in silico model built upon experimental data of single drugs and current systems biology information to

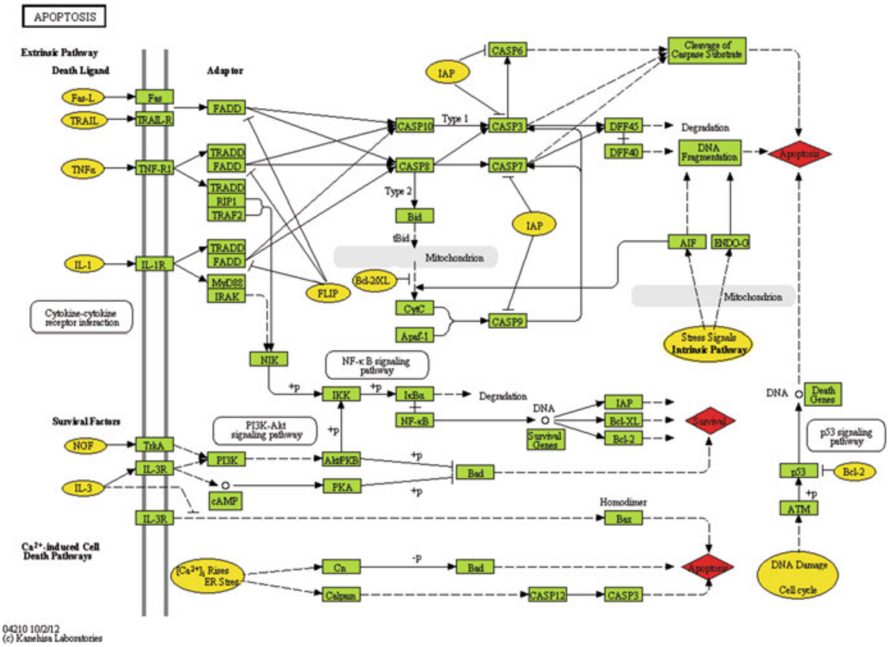


Fig. 3 The human apoptosis network extracted from the KEGG database (hsa04210). Genes are categorized as receptors (*yellow circles*), connecting genes (*green rectangles*), and the output nodes (*red diamonds*) that are implicated at the onset of the cell death machinery. A *solid line* with an *arrow* at the end indicates direct promotion; a *dashed line* with an *arrow* at the end indicates indirect promotion; a *dashed line* with a *bar* at the end indicates inhibition. A *cross symbol* between two genes indicates dissociation, in which case the two genes may be viewed as a single node (e.g., DFF45 and DFF40)

obtain maximum likelihood estimate (Fang et al. 2016). We briefly summarize the method below and discuss its potential applications in drug development process.

Biological networks are controlled/regulated by receptors. They are comprised of connecting genes and output nodes that are implicated in determining activation of the cell death machinery. Figure 3 presents a typical example network—apoptosis related signaling from the KEGG database (hsa04210). Different nodes have various signal propagation rules. Consider a combination study of s drugs $A_1, A_2, \ldots, A_s$, we first develop a statistical rescaling model to describe the effects of drugs on network topology. The model comprises a Hill equation for signals arriving at each receptor, a generic enzymatic rate equation to transmit signals among connecting genes, and a regression model to represent the cumulative effect of genes implicated in activation of the cell death machinery. Specifically, for a given dose-level $\boldsymbol{x} = (x_1, x_2, \ldots, x_s)^T$ of drugs $A_1, A_2, \ldots, A_s$, denote $a_{0i}(\boldsymbol{x})$ as the signal of receptor i obtained ($i = 1, 2, \ldots, r$) and $a_i(\boldsymbol{x})$ as the signal connecting gene i obtained ($i = 1, 2, \ldots$, r). Gene activity levels often exhibit a non-linear relationship to their upstream regulatory signals. Typically, a Hill equation (Weiss 1997) can be used to model the activity $a_{0i}(\boldsymbol{x})$ at receptor i,

$$a_{0i}(\boldsymbol{x}) = \frac{\left(\boldsymbol{\beta}_i^T\boldsymbol{x}\right)^{\alpha_i}}{1 + \left(\boldsymbol{\beta}_i^T\boldsymbol{x}\right)^{\alpha_i}}, \quad for \ \ i = 1, 2, \ldots, r, \tag{19}$$

where α_i and $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \cdots, \beta_{is})^T$ are the parameters to be estimated. To characterize the transmission of signals among connecting genes, the generic enzymatic rate equations can be used to adjust for possible feedback loops. Such equations have been motivated by various computational and biological considerations, a result of the close interaction between experimental and computational efforts (Lee et al. 2007; Ao et al. 2008). Let $a_j(\boldsymbol{x})$ be the activity at gene j and $a_{(i,j)}(\boldsymbol{x})$ the signal sending from gene j to gene i. The activity $a_i(\boldsymbol{x})$ at gene i is defined to be the summation of all signals $a_{(i,j)}(\boldsymbol{x})$ for gene j linked up gene i, and the generic enzymatic rate equation then suggests that

$$a_i(\boldsymbol{x}) = \Sigma_{j\in n(i)} a_{(i,j)}(\boldsymbol{x}), \text{ and } a_{(i,j)}(\boldsymbol{x}) = \frac{V_{F_j}\frac{a_j(\boldsymbol{x})}{\omega} - V_{B_i}\frac{a_i(\boldsymbol{x})}{\omega}}{\frac{V_{F_j}^2}{V_{F_j}^2+V_{B_i}^2}\left(1 + \frac{a_j(\boldsymbol{x})}{\omega}\right) + \frac{V_{B_i}^2}{V_{F_j}^2+V_{B_i}^2}\left(1 + \frac{a_i(\boldsymbol{x})}{\omega}\right)}, \tag{20}$$

where $n(i)$ is the set of genes that signal to gene i, and ω is the expected steady state parameter. V_{F_i} and V_{B_i} are the forward and backward parameters, respectively. When the action between genes i and j is irreversible in the backward direction, $V_{B_i} = 0$. The number of parameters V_{F_i} and V_{B_i} may become large if many connecting genes exist in the network. The forward and backward parameters V_{F_i} and V_{B_i} of connecting gene i may differ with those of connecting gene j $(i \neq j)$. Statistical variations typically occur when signals pass though the network because of link instability, stochastic noise inherent in the signal propagation rules, and/or chaos phenomena from the presence of loops. To model the network efficiently, it is reasonable to assume that V_{F_i} and V_{B_i} $(i = 1, 2, \ldots)$ are random effects that are independently and identically distributed (*i.i.d.*) normal random variables with mean μ_1 and variance σ_1^2.

A linear model is used to represent the cumulative effect of genes implicated at activation of the cell death machinery. For a given dose-level $\boldsymbol{x} = (x_1, x_2, \ldots, x_s)^T$ of drugs $A_1, A_2, \ldots, A_s$, let $Y(\boldsymbol{x})$ be the observed viability and $\mathbf{a}(\boldsymbol{x}) = \left(a_{i_1}(\boldsymbol{x}), \ldots, a_{i_h}(\boldsymbol{x})\right)^T$ be the vector of the activities at genes $i_1, \ldots, i_h$ which activate the output, then we have

$$Y_k(\boldsymbol{x}) = u_0 + \mathbf{a}(\boldsymbol{x})^T\boldsymbol{u} + \epsilon_k(\boldsymbol{x}), \tag{21}$$

where the subscript k is the k-th replication at dose-level $\boldsymbol{x} = (x_1, x_2, \ldots, x_s)^T$, the measurement error is assumed to be $\epsilon_k(\boldsymbol{x}) \sim N\left(0, (\sigma(\boldsymbol{x}))^2\right)$, and the standard deviation $\sigma(\boldsymbol{x})$ of the measurement error may depend on the dose-level $\boldsymbol{x} =$

$(x_1, x_2, \ldots, x_s)^T$. u_0 is the intercept parameter, and $\boldsymbol{u} = (u_1, \ldots, u_h)^T$ is the vector of regression parameters to be estimated. The positive parameter u_i indicates promotion by gene i; the negative parameter u_j indicates inhibition by gene j. Since $\boldsymbol{a}(\boldsymbol{x})$ in model (21) equals to zero when $\boldsymbol{x} = \mathbf{0}$, the intercept μ_0 should be 100 % cell viability if there is no drug intervention on the network.

In combination studies, the data from single drug experiments are available a priori. To make models (19), (20), and (21) identifiable and estimate the multivariable dose–response, one will need some additional data on the drug combinations. Fang et al. (2016) have shown that this can be achieved with several combinations with each drug at its individual IC_{50}. Based on the training data, the parameters in Eqs. (19), (20), and (21) can be estimated with the maximum likelihood approach. Let $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_r^T\right)^T$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r)^T$ and $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \omega, \boldsymbol{u}^T, \sigma_0^2, \mu_1, \sigma_1^2\right)^T$ be the vector of all parameters to be estimated. Suppose that there are n distinct inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, and k_i replications at each input $\boldsymbol{x}_i$, the corresponding output is Y_{ij} for $j = 1, 2, \ldots, k_i$; and $i = 1, 2, \ldots, n$. For given μ_1, σ_1^2 and a sample V_{F_i} and V_{B_i} (i =1, 2, ...) from the normal distribution $N(\mu_1, \sigma_1^2)$, the ECM algorithm (Meng and Rubin 1993) can be applied to obtain the maximum likelihood estimation of $\boldsymbol{\beta}, \boldsymbol{\alpha}, \omega, \boldsymbol{u}, \sigma_0^2$. Furthermore, for given $\boldsymbol{\beta}, \boldsymbol{\alpha}, \omega, \boldsymbol{u}, \sigma_0^2$, we can obtain n samples of V_{F_i} and V_{B_i} using Eq. (21) with n distinct inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. The estimation of μ_1 and σ_1^2 can then be obtained.

The dose–response surface of multidrug combinations is complex and difficult to estimate adequately. To get sufficient information of drug interactions, the functional ANOVA (Sobol 1993, 2001, 2003) is employed, which similar to functional principal component analysis. Without loss of generality, assume the dose-level of s drugs $A_1, A_2, \ldots, A_s$, $\boldsymbol{x} = (x_1, x_2, \cdots, x_s)^T \in [0, 1]^s$, and $y = g(\boldsymbol{x})$ is the corresponding dose response. Let $g_0 = \int_{[0,1]^s} g(\boldsymbol{x})\, d\boldsymbol{x}$ be the overall mean of $g(\boldsymbol{x})$. Then, there is a unique decomposition

$$g(\boldsymbol{x}) = g_0 + \sum_{i=1}^{s} g_i(x_i) + \sum_{i<j} g_{ij}\left(x_i, x_j\right) + \cdots + g_{1,2,\ldots,s}(x_1, x_2, \cdots, x_s), \qquad (22)$$

which satisfies $\int_0^1 g_{i_1,\ldots,i_u}(x_{i_1}, \ldots, x_{i_u})\, dx_{i_k} = 0$, for any $1 \le u \le s$ and $1 \le k \le u$; and the orthogonality $\int_{[0,1]^s} g_{i_1,\ldots,i_u}(x_{i_1}, \ldots, x_{i_u})\, g_{j_1,\ldots,j_v}\left(x_{j_1}, \ldots, x_{j_v}\right) dx_1 \cdots dx_s = 0$ if $(i_1, \ldots, i_u) \ne (j_1, \ldots, j_v)$. The total and partial variances can be defined by

$$D = \int_{[0,1]^s} g^2(\boldsymbol{x})\, d\boldsymbol{x} - g_0^2 \text{ and } D_{i_1,\ldots,i_k} = \int_{[0,1]^k} g^2_{i_1,\ldots,i_k}(x_{i_1}, \cdots, x_{i_k})\, dx_{i_1} \cdots dx_{i_k}, \qquad (23)$$

respectively. Denote $D = \sum_{k=1}^{s} \sum_{i_1<\cdots<i_k} D_{i_1,\ldots,i_k}$, the ratios

$$R_{i_1,\ldots,i_k} = D_{i_1,\ldots,i_k}/D, 1 \leq i_1 < \cdots < i_k \leq s, \tag{24}$$

are called *global sensitivity indices* (Sobol 1993, 2001, 2003). The integer k is called the order of the index. All $R_{i_1,\ldots,i_k}$'s are non-negative and their sum $\sum_{k=1}^{s} \sum_{i_1<\cdots,i_k} R_{i_1,\ldots,i_k} = 1$. The equality $R_{i_1,\ldots,i_k} = 0$ implies that $g_{i_1,\ldots,i_k} = 0$ and so the interaction of drugs $A_{i_1} \cdots A_{i_k}$ is not significant. Significance of the interaction of drugs $A_{i_1} \cdots A_{i_k}$ decreases with decreasing $R_{i_1,\ldots,i_k}$ values. Hence, the dose–response model can be reduced if we only retain the principal terms with the largest global sensitivity indices, an approach similar to principal component analysis. It is also expected that the number of terms in the dose–response functional ANOVA representation will be reduced significantly because the cumulative global sensitivity indices of the first few terms usually contribute a dominant portion (say, 80 %) of the total variation (Fang et al. 2006). To obtain the numerical values of the global sensitivity indices, the Quasi-Monte Carlo methods for approximating the integrals can be adopted. For more details, please refer to Fang et al. (2006).

We have performed two simulation experiments to investigate the effectiveness of the optimal network simulator for the discovery of multidrug interactions using the apoptosis signaling network. The first example involves a combination study of five drugs; the second example considers as many as ten drugs. The simulation with five drugs identified three terms (drugs and their interactions) making most significant contributions yielding a global sensitivity indices of 90.45 %, which is consistent with the global sensitivity indices from the true dose–response. For the simulation with 10 drugs, it has been shown that the method identified four most significant terms with a total of global sensitivity indices of 92.19 %, which is consistent with the global sensitivity indices from the true dose–response. We summarize the process of the development of drug combinations in Fig. 4.

6 Discussions and Conclusions

Cancer cells carry out their functions following appropriate responses to the extracellular and intracellular inputs to their complex network of multiple signaling pathways. Many genes that code for proteins in these pathways are controlled by regulatory proteins that up-regulate or downregulate these genes depending on the inputs to the signaling network. It is conceivable that multidrug approach can play an even greater role in cancer developmental therapeutics with the development of the systems biology. For combinations of two and three drugs, we have reviewed the MTDs and statistical modeling of joint effects for experiments performed following the MPD. For phase I clinical trial, we outlined an approach that models the drug

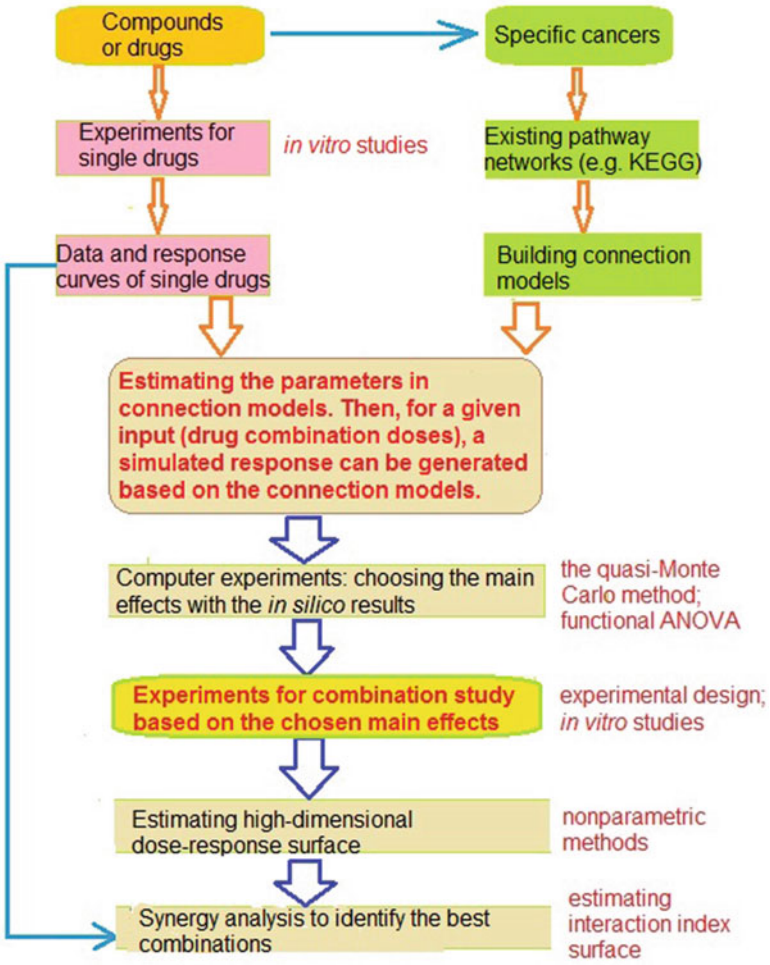

Compounds or drugs
Specific cancers
Experiments for single drugs
in vitro studies
Existing pathway networks (e.g. KEGG)
Data and response curves of single drugs
Building connection models
Estimating the parameters in connection models. Then, for a given input (drug combination doses), a simulated response can be generated based on the connection models.
Computer experiments: choosing the main effects with the in silico results
the quasi-Monte Carlo method; functional ANOVA
Experiments for combination study based on the chosen main effects
experimental design; in vitro studies
Estimating high-dimensional dose-response surface
nonparametric methods
Synergy analysis to identify the best combinations
estimating interaction index surface

Fig. 4 The framework for multiple drug combination study

interactions and the dose escalation algorithm based on a Bayesian model with computation performed using Markov chain Monte Carlo. However, combination studies are extremely difficult to implement in clinical trials suggesting the greater importance of preclinical testing of the combination based on cell lines and animal models that are well established.

For combinations of more than three drugs, we proposed a two-stage approach with the first stage utilizing data of single drugs (and some drug combinations) and the current network information to develop a statistical model to describe the drug effects on the network. Through these statistical models, we conducted computer experiments (in silico) to derive a global sensitivity index of each term in the functional ANOVA of dose response model by generating doses of the drugs with the Quasi Monte-Carlo method. Then, we can predict the main effects that occur with combinations of multiple drugs. Two simulation studies illustrate the superior performance of our methods. The principal global sensitivity indices generally select 3–4 terms of multidrug combinations in the functional ANOVA model if the true dose response function is smooth. Then we can develop experimental designs and statistical procedures on the few selected terms. Given the scope of this article, we will report the details of experimental design and data analysis based on these selected interaction and main effect terms in a future report.

Acknowledgements The research of Drs. Fang, Huang and Tan is partially supported by the National Cancer Institute (NCI) grant R01CA164717. Dr Yang's research was completed and partially supported by University of Maryland while she was a PhD student there.

References

Ao, P. Lee, L.W. Lidstrom, M.E. Yin, L. and Zhu, X. Towards kinetic modeling of global metabolic networks: Methylobacterium extorquens AMI growth as validation. *Chinese Journal of Biotechnology* **24**: 980–994, 2008.

Ashton, J.C. ANOVA and the analysis of drug combination experiments. *Nature Methods* **12**: 1108, 2015.

Berenbaum, M. C. What is Synergy? *Pharmcological Reviews* **41**: 93–141, 1989.

Bliss, C.I. The toxicity of poison applied jointly. *Annals of Applied Biology* **18**: 585–815, 1939.

Calzolari, D. *et al*. Search algorithms as a framework for the optimization of drug combinations. *PLoS Computational Biology* **4**(12): e1000249, 2008.

Chou, T.C. and Talalay, P. Quantitative analysis of dose–effect relationships: the combined effects of multiple drugs or enzyme inhibitors. *Advances in Enzyme Regulation* **22**: 27–55, 1984.

Fang, H.B. Chen, X. Pei, X.Y. Grant, S. and Tan, M. Experimental Design and Statistical Analysis for Three-Drug Combination Studies. Statistical Methods in Medical Research, 2015. DOI: 10.1177/0962280215574320.

Fang, H.B. Huang, H. Clarke, R. and Tan, M. Predicting multi-drug inhibition interactions based on signaling networks and single drug dose–response information. *Journal of Computational Systems Biology*, 2, 2016.

Fang, H.B. Ross, D.D. Sausville, E. and Tan, M. Experimental design and interaction analysis of combination studies of drugs with log-linear dose responses. Statistics in Medicine, **27**(16):3071–3083, 2008.

Fang, K.T. Li, R. and Sudjianto, A. *Design and Modeling for Computer Experiments*. Chapman & Hall/CRC: New York, 2006.

Fang, K.T. and Wang, Y. *Number-Theoretic Methods in Statistics*. London: Chapman and Hall, 1994.

Fitzgerald, J.B. Schoeberl, B. Nielsen, U.B. and Sorger, P.K. Systems biology and combination therapy in the quest for clinical efficacy. *Nature Chemical Biology* **2**: 458–466, 2006.

Gojo, I. Tan, M. Fang, H.B. Sadowska, M. Lapidus, R. Baer, M.R. Carrier, F. Beumer, J.H. Anyang, B.N. Srivastava, R.K. Espinoza-Delgado, I. and Ross, D.D. Translational phase I trial of Vorinostat (suberoylanilid.e hydroxamic acid) combined with cytarabine and etoposide in patients with relapsed, refractory, or high-risk acute myeloid leukemia, *Clinical Cancer Research* **19**:1838–1851, 2013.

Goldoni, M. and Johansson, C. A mathematical approach to study combined effects of toxicants in vitro: Evaluation of the Bliss independence criterion and the Loewe additivity model. *Toxicology in Vitro* **21**: 759–769, 2007.

Greco, W.R. Bravo, G. and Parsons, J. C. The Search for Synergy: A Critical Review from a Response Surface Perspective. *Pharmcological Reviews* **47**: 331–385, 1995.

Hopkins, A.L. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology* **4**(11): 682–690, 2008.

Krzywinski, M. and Altman, N. Two-factor designs. *Nature Methods* **11**: 1187–1188, 2014.

Lee, L.W. *et al*. Generic enzymatic rate equation under living conditions. *Journal of Biological Systems* **15**: 495–514, 2007.

Loewe, S. Isobols of Dose-Effect Relations in the Combination of Pentylenetetrazole and Phenobarbital. *Journal of Pharmacology and Experimental Therapeutics* **114**: 185–191, 1955.

Meng, X.L. and Rubin, D.B. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**: 267–278, 1993.

Pei, X.Y. Dai, Y. and Grant, S. The proteasome inhibitor bortezomib promotes mitochondrial injury and apoptosis induced by the small molecule Bcl-2 inhibitor HA14-1 in multiple myeloma cells. *Leukemia* **17**: 2036–2045, 2003.

Pei, X.Y. Dai, Y. and Grant, S. The small-molecule Bcl-2 inhibitor HA14-1 interacts synergistically with flavopiridol to induce mitochondrial injury and apoptosis in human myeloma cells through a free radical-dependent and Jun NH2-terminal kinase-dependent mechanism. *Molecular Cancer Therapeutics* **3**: 1513–1524, 2004.

Peterson, J.J. and Novick, S. J. Nonlinear Blending: A Useful General Concept for the Assessment of Combination Drug Synergy, *Journal of Receptors and Signal Transduction* **27**(2): 125 – 146, 2007.

Shiozawa, K., Nakanishi, T., Tan, M., Fang, H.B., Wang, W.C., Edelman, M.J., et al. Preclinical studies of Vorinostat (suberoylanilide hydroxamic acid) combined with cytosine arabinoside and etoposide for treatment of acute leukemias. *Clinical Cancer Research* **15**:1698–1707, 2009.

Sobol, I.M. Sensitivity analysis for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment* **1**: 407–414, 1993.

Sobol, I.M. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. & Comp. in Simulation* **55**: 271–280, 2001.

Sobol, IM. Theorems and examples on high dimensional model representation. *Reliability Engineering & System Safety* **79**: 187–193, 2003.

Sun, Z. and Braun, T.M. A two-dimensional biased coin design for dual-agent dose-finding trials. *Clinical Trials*, doi: 10.1177/1740774515592404, 2015.

Tan, M. Fang, H.B. and Tian, G.L. Dose and sample size determination for multi-drug combination studies. *Statistics in Biopharmaceutical Research* **1**: 301–316, 2009.

Tan, M. Fang, H.B. Tian, G.L. and Houghton, P.J. Experimental design and sample size determination for drug combination studies based on uniform measures. *Statistics in Medicine* **22**: 2091–2100, 2003.

Tian, G.L. Fang, H.B. Tan, M. Qin, H. and Tang, M.L. Uniform distributions in a class of convex polyhedrons with applications to drug combination studies. *Journal of Multivariate Analysis* **100**: 1854–1865, 2009.

Wages, N.A., Conaway M.R., and O'Quigley J. Continual reassessment method for partial ordering. *Biometrics* **67**(4): 1555–63, 2011.

Weiss, J.N. The hill equation revised: uses and misuses. *FASEB journal* **11**: 835–841, 1997.

Xavier, J.B. and Sander, C. Principle of System Balance for Drug Interactions. *The New England Journal of Medicine* **362**(14): 1339–1340, 2010.

Yang, Y., Fang, H.B., Roy, A. and Tan, M. An adaptive Bayesian dose finding approach for drug combinations with drug-drug interaction. *Statistics and Its Interface* (in review), 2016.

Yin, G. and Yuan, Y. A latent contingency table approach to dose finding for combinations of two agents. *Biometrics* **65**(3): 866–875, 2009a.

Yin, G. and Yuan, Y. Bayesian dose finding in oncology for drug combinations by copula regression. *Journal of the Royal Statistical Society: Series C* (Applied Statistics) **58**(2): 211–224, 2009b.

Yuan, Y. and Yin, G. Sequential continual reassessment method for two-dimensional dose finding. *Statistics in Medicine* **27**(27): 5664–5678, 2008.

Statistical Methods for Analytical Comparability

Leslie Sidor

Abstract In all manufacturing settings, there is an inherent drive to improve product through the reduction in process variation, implementing new technology, increasing efficiency, optimizing resources, and improving customer experience through innovation. In the pharmaceutical industry, these improvements come with added responsibility to the patient such that product made under the post-improvement or post-change condition maintains the safety and efficacy of the pre-change product. Regulatory agencies recognize the importance in providing manufacturers the flexibility to improve their manufacturing processes (FDA, Guidance Concerning Demonstration of Comparability of Human Biological Products, 1996; ICH Q5E, ICH Guidance for Industry: Q5E Comparability of Biotechnology/Biological Product Subject to Changes in Their Manufacturing Process, 2005). They also acknowledge that some changes may not require additional clinical studies to demonstrate safety and efficacy so that implementation may be more efficient and expeditious to benefit patients. When clinical studies are not necessary, a minimum requirement remains to demonstrate that the post-change product is comparable to the pre-change product. This comparison is known as analytical comparability. Analytical comparability may be demonstrated through the use of statistical and non-statistical methods. The choice of the methodology is not defined by the guidance documents. This paper presents an overview and use of equivalence tests and statistical intervals as options to demonstrate analytical comparability.

Keywords Analytical Comparability • Equivalence Tests • Equivalence Acceptance Criteria • Tolerance Intervals

L. Sidor (✉)
Biogen, 300 Binney Street, Cambridge, MA 02142, USA
e-mail: Leslie.Sidor@Biogen.com

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_19

1 Introduction

Across the regulatory documents such as ICH Q5E (2005) and FDA guidance (1996), there are only high level recommendations for the design of a comparability study and for setting acceptance criteria to assess the impact of the change on the product. There is no recommendation on the approach for setting acceptance criteria or the statistical techniques that should be used to compare pre- and post-change data. This paper will provide an overview of methodologies for setting analytical comparability acceptance criteria, design considerations and the statistical techniques that are available to the scientist as it relates to analytical comparability.

The typical categories of statistical tests and methods to establish acceptance criteria for analytical comparability are summarized in Table 1. The approaches used to demonstrate and establish acceptance criteria can be categorized using a recommendation described in the article by Chatfield et al. (2011). In particular, methods to demonstrate comparability are categorized as either *equivalence tests* or as *other comparability approaches*. Regardless of the approach selected from Table 1, the acceptance criteria need to be pre-defined. For convenience, the word "comparability" will be used in future text to describe the comparison of a pre-and post-change process in an analytical comparability study and the words "product" and "process" may be used interchangeably except where distinct differences are called out.

The most important factor in selecting a comparability approach is to select the approach that best matches the desired definition of comparability. If comparability is more logically defined in terms of estimable process and product parameters (e.g. means and slopes), the preferred approach is an equivalence test. Three examples of this situation are:

1. Comparisons where it is desirable to have comparable means,
2. A situation where a shift in the mean between two conditions is important in consideration of meeting specifications, and
3. Profile studies where it is desirable to have comparable slopes.

Table 1 A summary of the approaches and comparability categories typically utilized in analytical comparability

Acceptance criteria approach	Lot release and in-process parameters	Stability at recommended storage conditions	Stability at stressed storage conditions	Characterization methods
Equivalence tests	X		X	X
Specifications	X	X		
Tolerance intervals	X			X
Prediction intervals	X			X
Visual comparisons	X	X	X	X
Limit evaluations	X	X		X

Conversely, if the comparison of population means or slopes is not of interest, then a comparability approach based on one of the other approaches may be more reasonable. Of the approaches listed in Table 1, an equivalence test is the only approach that uses a formal statistical test that can be used to explicitly control both the manufacturer's risk and consumer's risk.

Once the methodology has been selected, the comparability study is designed. The study design needs to align with the selected acceptance criteria approach. Since the equivalence test is the only statistical test listed in Table 1, the concepts associated with hypothesis testing are used to select the study design and will consider both the manufacturer's (type I error) and consumer's (type II error) risk. When a tolerance or prediction interval is used for comparability, the study design simply consists of prescribing the number of post-change lots that must fall within the acceptance criterion.

2 Data Overview

One of the key items to consider when designing a comparability study and the acceptance criteria is the source and structure of the data. This understanding is done collaboratively with the statistician and the scientist. It is important for the statistician to understand how data are collected, the availability of data, analytical method changes, limitations associated with randomizing samples on laboratory equipment, etc. Also, the scientist needs to be aware of limitations in the data such as a large proportion of the data below the limit of detection, data that are stratified due to rounding practices, and impact of analytical method changes on data, etc.

2.1 Data Types

In general, the data types that are typical of comparability studies can be organized into two categories: continuous and discrete. Within these categories, data may be collected with replication; there may be multiple scales; and the data may be collected over time. The types of data and the likely data source and situation are summarized in Table 2.

Understanding the different data types prior to setting the acceptance criteria enhances the collaborative discussion with the subject matter expert to outline the risks and constraints associated with each comparability approach.

Table 2 Summary of data types by data source most likely data source

Data type	Data source	Likely situation
Continuous—no replication	Lot release, characterization assays	Purity, Potency, pH, etc.
Continuous—with replication	Lot release, characterization assays	Fill weight, orthogonal bioassays
Continuous—with values < Limit of Detection or Limit of Quantitation	In-process specification, characterization assays	Impurities, clearance assays
Continuous—with multiple scales or multiple reference standards	Lot release, in-process specification, characterization assays	Protein concentration (clinical scale to commercial scale)
Discrete	Lot release, characterization assays	Particle counts, sub-visible particle counts
Continuous longitudinal—no replication	Stability at recommended storage conditions or stressed conditions	Purity, potency etc.
Continuous longitudinal—replication	Stability at stressed conditions	Purity, potency etc.

3 Statistical Overview

3.1 Tolerance and Prediction Intervals

As a comparability approach, prediction and tolerance intervals are appropriate for evaluating individual values versus a statistical parameter such as a mean. When using a prediction or tolerance interval to assess comparability, the definition of "passing" consists of a pre-specified number of post-change observations falling within the interval. Therefore, the width of the interval is a key component in setting these types of acceptance criteria. The width of the interval is dependent on the level of confidence, coverage and sample size. Using a tolerance interval with a small data set will generally produce an interval that may be too wide to be helpful in assessing comparability. In these situations, a prediction interval may be more useful since the width of the interval will be partly controlled by the number of future observations that are desired to fall within the interval. Regardless of the interval choice, the basic statistical assumptions apply and need to be evaluated, and the final interval must fall within the specification limit.

A tolerance interval is designed to contain $P \times 100\%$ of all future observations with $(1 - \alpha)$ 100% confidence. The basic form of a two-sided tolerance interval is

$$TI = \overline{Y} \pm K\sqrt{S^2} \tag{1}$$

where $\overline{Y}$ is the sample mean, K is a constant dependent on the sample size and coverage, and S^2 is the variance. A prediction interval differs from a tolerance interval in that it is designed to capture a specified number of future observations with $(1 - \alpha)$ 100% confidence. The basic form of a two-sided prediction interval is

$$PI = \overline{Y} \pm t_{1-\alpha/2m;n-1}\sqrt{\left(1 + \frac{1}{n}\right) \times S^2} \tag{2}$$

where $\overline{Y}$ is the sample mean, $t_{1-\alpha/2m;n-1}$ is the t statistic for m future post-change observations with $n - 1$ degrees of freedom, and n is the number of observations used to compute $\overline{Y}$. The computation of the terms will change based on the data structure. Although a prediction intervals will not be computed for the example, the computation of $\overline{Y}$ and S^2 will still apply.

Once the intervals are computed, the post-change data are collected. Since the use of intervals are not a statistical test, the study is not designed to control Type I or Type II errors. To “pass” comparability, the predefined number of post-change lots simply need to fall within the computed interval.

3.2 Tolerance and Prediction Intervals

The first example uses independent measures at two manufacturing scales. This scenario is typical when the attribute of interest is dependent on manufacturing scale or a known and acceptable bias between two laboratories. In this example, a tolerance interval is needed to transfer a commercial scale process to another commercial site. The parameter of interest is an in-process attribute, protein yield in kilograms. The specification for the commercial scale process is 40.8–69.0 kg. The data available to compute the tolerance interval are five lots from a smaller clinical scale and five lots from the current commercial scale process. The data from the clinical scale and the current commercial scale are plotted in Fig. 1 and listed in the Appendix.

For the acceptance criterion, a two-sided tolerance interval is chosen where $P = 0.99$ and $\alpha = 0.05$ which implies that the interval will contain 99 % of the population with 95 % confidence. By combining the data, the multiplier K will be smaller thus reducing the width of the interval. The decrease in K in this simple example is due to the chi-squared value increasing with an increase in the degrees of freedom. With a sample size of ten lots, K is reduced from 6.60 to 4.8283. Refer to Eq. (4) for the computation of K. Calculation of the tolerance interval is more complex due to the fixed effect, scale, and the need to center the interval on the commercial scale mean. The independent model in Eq. (3) is used to estimate the variance and mean. Because there are two process scales, an additional subscript is required where Y_{ij} is the measured value for lot j collected from scale i, n_i is the number of lots in scale i ($n_1 = n_2 = 5$ in this example), μ_1 and μ_2 are the true

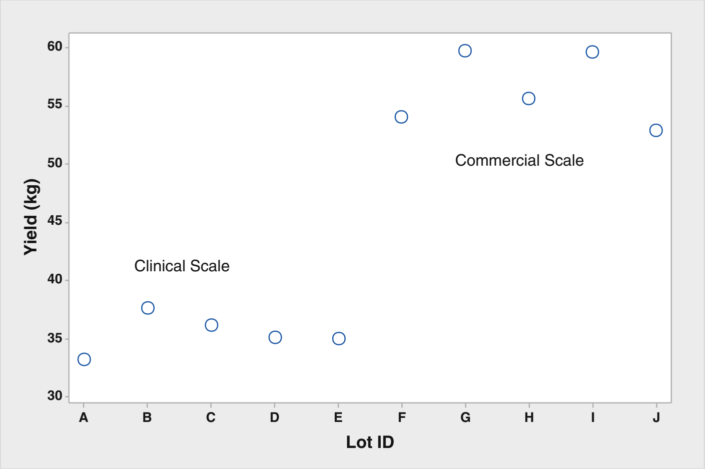


Fig. 1 Pre-change clinical and commercial scale data plotted in time order

Table 3 Values required to compute a tolerance interval with a single fixed effect and common variance

Statistic description	Notation	Formula	Values
Predicted value of Y at the center of interest	$\overline{Y}_i$		56.365
Root mean squared error	RMSE	$\sqrt{S^2}$	2.49247
Error degrees of freedom	Error df	$n_1 + n_2 - 2$	8
Effective sample size	n_e	n_i	5
Confidence level	$(1-\alpha)$	NA	0.95
Proportion contained	P	NA	0.99

averages for the clinical and commercial scales respectively, and E_{ij} is the residual error associate with lot j and scale i. It is assumed that the E_{ij} are sampled from a normal population with mean 0 and variance σ_E^2.

$$\begin{aligned} &Y_{ij} = \mu_i + E_{ij} \\ &i = 1,\ 2\ \text{(scale)};\ \ j = 1, \ldots, n_i\ \text{(lots per scale)} \end{aligned} \tag{3}$$

To compute this interval, the values listed in Table 3 are required.

The Minitab output can be used to obtain the predicted mean for the commercial scale process, the scale of interest and the root mean square error. These values are taken from Table 4. The root mean square is obtained by taking the square root of the Mean Square Error term, $\sqrt{6.21} = 2.492$, in the Analysis of Variance table below (Table 4).

Table 4 Analysis of variance for Example 1

Source	DF	Adj SS	Adj MS	F-Value	p-value
Scale	1	1100.54	1100.54	177.15	0.000
Error	8	49.70	6.21		
Total	9	1150.24			

To compute the tolerance interval, Eq. (1) is used where K is computed using an approximation (Howe 1969) shown in Eq. (4) and S^2 is the pooled error across both manufacturing scales. For the example, K is

$$K = \sqrt{\frac{\left(1 + \frac{1}{n_e}\right) Z^2_{(1+P)/2} \times (\text{error df})}{\chi^2_{\text{error df},\alpha}}} = \sqrt{\frac{\left(1 + \frac{1}{5}\right) 2.576^2 \times 8}{2.7326}} = 4.8283 \tag{4}$$

and, the lower and upper bounds of the tolerance interval are

$$\begin{aligned} L &= 56.360 - 4.8283 \times 2.49247 = 44.3256 \\ U &= 56.360 + 4.8283 \times 2.49247 = 68.3944 \end{aligned} \tag{5}$$

The tolerance interval containing 99 % of all future observation with 95 % confidence is 44.326–68.394 kg which falls within the in-process specification. If this interval were to be used for an acceptance criterion, the upper limit is so close to the upper specification limit of 69.0 kg, it may make more sense to default to the specification for the upper limit only. One could also consider choosing an interval with less coverage or consider using a prediction interval designed to predict the next m number of post-change lots. Note that if only the five commercial lots were used to compute the tolerance interval, the resulting interval would be 35.61–77.17 kg which exceeds the lower and upper specification. Therefore, incorporating the clinical lots to increase the error df which in return increased the chi-squared statistic thus reducing the multiplier K to narrow the interval. In this example, the additional clinical lots also reduced the variance. This may not always be the case. Figure 2 provides a plot of the pre-change commercial scale data, the commercial scale specification and the two sets of two-sided 95 % confidence/99 % coverage tolerance intervals. Plots like Fig. 2 aid in the visualization of the acceptance criterion relative to the data and the specification.

Once the post-change data are collected, comparability "passes" if the post-change data fall within the interval. In addition to the pass/fail assessment, these data should be plotted in time order along with the pre-change data to provide a visual assessment of potential step shifts in the process.

Although the example presented represents a simple case, tolerance and prediction intervals can be computed given more complex data structures (e.g. replicate measurements taken per lot (balanced and unbalanced), count data, and censored data). The distributions and the corresponding models need to align with the particular data structure. For a more detailed discussion on setting intervals for some complex data structures, the reader is referred to Hahn and Meeker (1991).

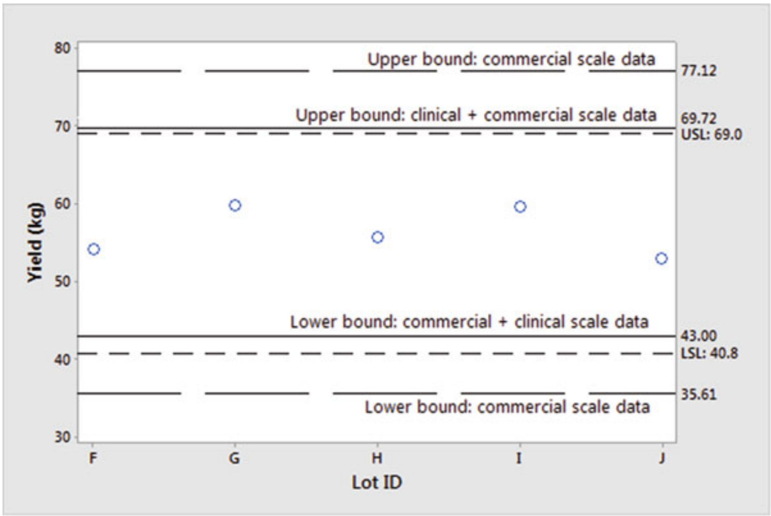


Fig. 2 Commercial scale pre-change data by lot ID with acceptance criteria and specification

3.3 Equivalence Tests

As noted previously, the use of an equivalence test to assess comparability is the only statistical test listed in Table 1. Data may be collected across multiple lots at a single point in time such as lot release. These data are referred to as non-profile data. For non-profile data, the comparison of interest is the difference in the pre and post-change means. Data may also be collected over time across multiple lots and multiple time points such as a stressed stability study. These data are referred to as profile data, and the comparison of interest is the difference in the pre and post-change slopes. Because both comparisons are statistical tests, the comparison can be defined in terms of a set of hypotheses. The hypotheses are of the form for non-profile data in Eq. (6) and for profile data in Eq. (7)

$$\begin{aligned} H_0 &: \left|\mu_{pre} - \mu_{post}\right| \geq \mathrm{EAC} \\ H_a &: \left|\mu_{pre} - \mu_{post}\right| < \mathrm{EAC} \end{aligned} \tag{6}$$

$$\begin{aligned} H_0 &: \left|\gamma_{pre} - \gamma_{post}\right| \geq EAC \\ H_a &: \left|\gamma_{pre} - \gamma_{post}\right| < EAC \end{aligned} \tag{7}$$

where EAC is the acronym for equivalence acceptance criterion, and the subscripts on μ and γ represent the product means or slopes for the pre-and post-change product. A test of equivalence is assessed by constructing a two-sided $100\,(1-2\alpha)\,\%$ confidence interval on the difference $\mu_{pre} - \mu_{post}$ or $\gamma_{pre} - \gamma_{post}$. The null hypothesis H_0 in Eqs. (6) and (7) is rejected, and equivalence is demonstrated if the entire confidence interval falls in the range of $-$EAC to $+$ EAC. This procedure provides

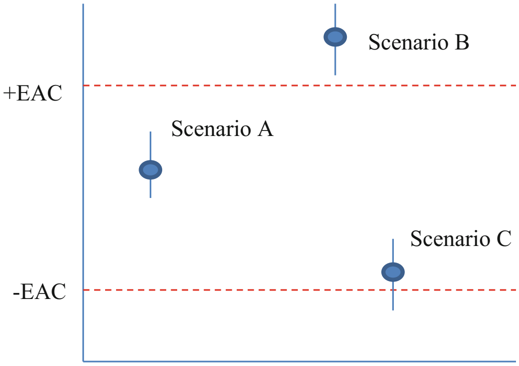


Fig. 3 Outcomes of an equivalence test

a test size of less than or equal to 100α %. In the literature, alpha is typically set at 0.05 (Limentami et al. 2005; Schuirmann 1987). The Type II error, β, is dependent on the study design. Ideally, the power of the study, $(1 - \beta)$, should be at 0.99 when the true difference in parameters is zero. If adequate power cannot be achieved, the risks should be discussed with the subject matter expert. Failing to reject H_0 does not imply that the pre-and post-change products are *not* equivalent.

The results of an equivalence test can be described graphically. Figure 3 presents three possible outcomes of an equivalence test. Scenario A is the situation where the confidence interval is entirely contained in the range from –EAC to +EAC. The conclusion for Scenario A is to reject H_0 and claim that the pre-and post-change process means are equivalent. Scenarios B and C reflect the situation where H_0 is not rejected. Both scenarios have potentially different consequences in the context of demonstrating equivalence. For Scenario B, the entire confidence interval falls above +EAC. This implies that the two product means are not equivalent which is a more serious consequence of failing to reject H_0. Scenario C is considered inconclusive (Chatfield and Borman 2009). There is not enough evidence to declare equivalence or nonequivalence. This situation is typically observed when the post-change variance is large relative to the pre-change variance or the power of the initial study design was too low.

The most critical element of the equivalence test is the choice of EAC. When evaluating process or product comparability, the EAC defines the maximum difference in means that has no practical scientific impact. It is ideal if the subject matter expert can define the EAC. An EAC may also be defined by specifications or other decision making limits where the EAC is based on maintaining a high probability of meeting these limits for a given process shift. In the case where there is no specification or no subject matter expert guidance to define a meaningful shift as it relates to safety and efficacy, statistics may be used to compute a preliminary EAC. A statistical EAC describes the difference in means based on "expected" behavior of the pre-change process versus "acceptable" in terms of safety or efficacy. The notion of "expected" behavior is proposed by Hauk et al. (2008). The historical behavior can include lot to lot variation along with the intermediate precision of the analytical method. Note that the use of this technique is in the absence of a scientific definition of practical significance and is typically used when computing an EAC for characterization methods or stressed stability studies. When computing a statistically derived EAC, a statistical model is required to describe the historical behavior of the pre-change process.

A way to define a statistical EAC is to use the definition of an effect size. When defining the EAC as an effect size (ES), the acceptable shift in population means or slopes is expressed in terms of the standard deviation of the response variance. Therefore, the important terms in defining the statistical EAC are the variance estimates. Mathematically, the ES can be defined as Eq. (8) for non-profile data as

$$\mathrm{ES} = \frac{|\mu_{pre} - \mu_{post}|}{\sqrt{\sigma_E^2}} \tag{8}$$

where σ_E^2 is the variance associated with the normal population of the E_{ij}. This variance estimate is used when an independent model is fit. The denominator in Eq. (8) may be modified to include additional variance components such as a random lot effect if necessary.

Solving for $|\mu_{pre} - \mu_{post}|$, the EAC becomes

$$\mathrm{EAC} = \pm \mathrm{ES}\sqrt{S_E^2} \tag{9}$$

For profile data, the statistical model used is the random intercept model. The assumed model for establishing the preliminary EAC using the pre-change data is

$$\begin{aligned} Y_{ij} &= \mu + L_i + \gamma \times t_{ij} + E_{ij} \\ i &= 1, \ldots, n;\ j = 1, \ldots, T_i \end{aligned} \tag{10}$$

where Y_{ij} is a response measured for lot i at time point j, μ is the average y-intercept across all lots, γ is the average slope across all lots, L_i is a random variable that allows the y-intercept to vary from μ for a given lot, L_i has a normal distribution with mean 0 and variance σ_L^2, t_{ij} is the time point for measurement j of lot i, E_{ij} is a random normal error term created by measurement error and model misspecification with mean 0 and variance σ_E^2, n is the number of sampled lots, T_i is the number of time points obtained in lot i, and L_i and E_{ij} are jointly independent. Like in the non-profile case, the variance is partitioned into the lot-to-lot variance component and the random error which represents the measurement error and model misspecification. Since it is assumed that the reaction kinetics driving the stability properties are consistent between the pre-and post-change processes, the random slope effect is not fit in the model. The effect size is

$$\mathrm{ES} = \frac{|\gamma_{pre} - \gamma_{post}|}{\sqrt{Var\left(\widehat{\gamma}_{pre}\right)}} \tag{11}$$

where $Var\left(\widehat{\gamma}\right)$ is variance of the slope estimate and is defined as

$$Var\left(\widehat{\gamma}\right) = \frac{S_E^2}{SST} \tag{12}$$

where $SST = \sum_{j=1}^{T} \left(t_j - \bar{t}\right)^2$ and $\bar{t} = \sum_{j=1}^{T} t_j \Big/ T$. When computing the SST, only unique time points are used. The EAC is

$$\text{EAC} = \text{ES}\sqrt{\frac{S_E^2}{SST}} \tag{13}$$

Regardless of the data type, choosing the effect size may be challenging for the subject matter expert. In these situations, it is helpful to show the effect size graphically. For non-profile data, the consequence of different effect sizes may be shown graphically as overlapping normal curves. In Fig. 4, four effect sizes are presented. The pre-change population is represented by a dashed line and the post-change population is represented by a solid line. In the first panel, top-left, the effect size is zero which corresponds to 100 % overlap between the pre and post-change populations. As the effect size increases, the amount of overlap decreases. The most extreme case presented in Fig. 4 is an effect size of three. In this situation there is only a 13 % overlap between the pre and post-change populations (Lei and Olson

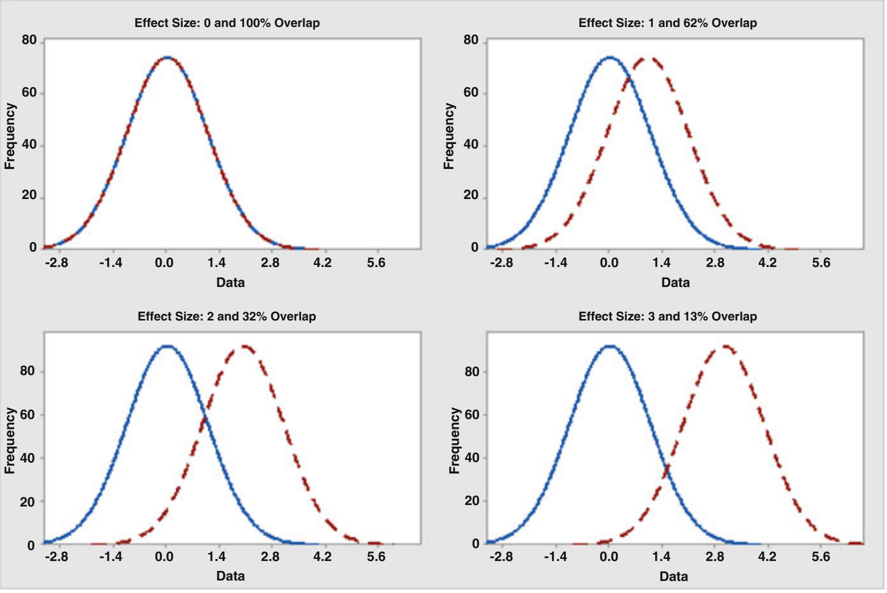


Fig. 4 Plots of effect sizes (ES) with overlapping pre-change (*dashed line*) and post-change (*solid line*) populations with percent population overlap

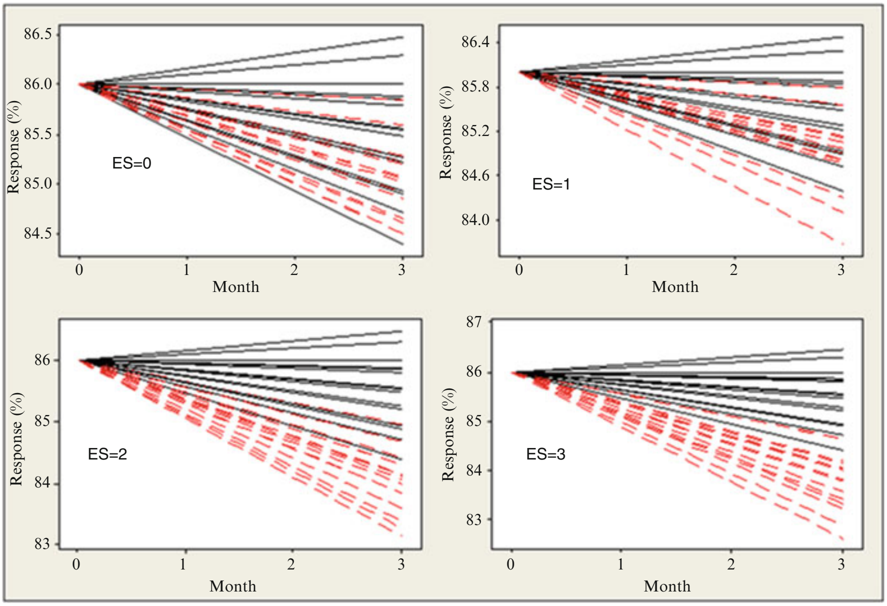


Fig. 5 Plots of effect sizes (ES) with stability profiles with x-axis label of month and y-axis label of response (%) (*solid lines* represent the pre-change process and dashed lines represent the post-change process)

2010). Another important consideration when using these plots is that the overlap is associated with the most extreme difference of the pre-change and post-change means. When the mean shift is equal to the ES in each panel of Fig. 4 there is only a 5 % chance that the difference in the two means would pass the equivalence test.

The same visual concept is presented graphically in Fig. 5 to evaluate the percent overlap of degradation profiles when evaluating an effect size for a stressed stability EAC. As noted above, the EAC for a stressed stability study involves the comparison of the pre and post-change slopes.

After the EAC have been established, the study is designed. Because an equivalence test is a statistical test, the Type I and Type II errors are defined. As noted previously, convention is to fix the Type I error to 0.05 which controls the risk to patient. The manufacturer's risk or the Type II error is controlled through an adequate study design. The more data one collects, the better the power of the study. An added benefit of collecting more data is that one gains additional understanding of the post-change process. Depending on the type of model fit, improving the power may be accomplished in a couple of ways. For non-profile data with no random lot effect, power may be increased by increasing the number of replicates per lot. If there is a significant lot effect, the number of lots tested needs to be increased and increasing the number of replicates will have little impact on improving the power of the study. For profile data, the same concepts apply. However, there is an added complexity of choosing the time points for the study. In general, replicating time points at the beginning and the end of the stressed stability study will have a greater impact on reducing the variance of the slope thus improving power. If replicating time points is not possible, one could consider two to three time points closely spaced at the beginning and end of the stressed stability study.

Once the post-change data are collected, the equivalence test is conducted. To test the hypothesis in Eqs. (6) or (7), a one-sided 95 % lower bound on the difference in parameters and a one-sided 95 % upper bound on the difference in parameters is constructed. Together, these bounds form a 90 % two-sided confidence interval on $\mu_{pre} - \mu_{post}$ or $\gamma_{pre} - \gamma_{post}$. The null hypothesis, H_0, is rejected and equivalence is demonstrated if the computed interval falls in the range from −EAC to +EAC. This test provides a Type I error rate of 5 %. In other words, if the two slopes or means differ in absolute value by greater than the EAC, there is no greater than a 5 % chance of declaring the two processes are equivalent (i.e. rejecting H_0).

3.4 Equivalence Test: Non-Profile Data

For this example, an EAC is computed using a specification as a control for an acceptable shift in process means. The data are final protein concentration collected as a lot release parameter. The specification for protein concentration is 65.0 mg/mL ± 10 % or 58.5–71.5 mg/mL. The historical process data are plotted in Fig. 6, and the descriptive statistics are listed in Table 5.

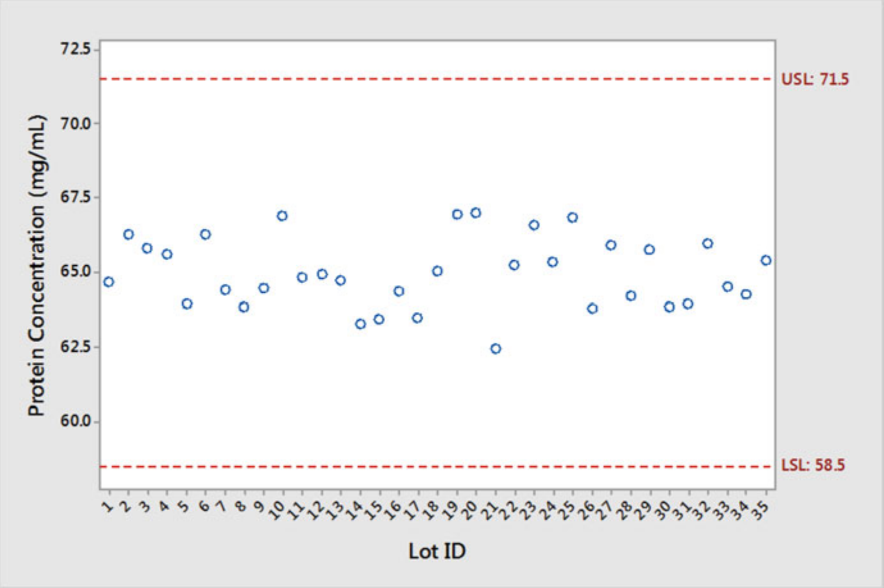


Fig. 6 Individual value plot of protein concentration (mg/mL) by lot

Table 5 Descriptive statistics for non-profile data (protein concentration)

Statistic	Value (mg/mL)
Mean	65.003
Standard deviation	1.179

To compute an EAC given a specification limit, one can ask the question, "Given the pre-change process mean is 65.003 mg/mL and the standard deviation of 1.179 mg/mL, what is the maximum allowable shift in the post-change mean that would not cause an unacceptable probability for an out of specification observation?" This question can be answered by using a process capability index. First, consider the capability of the current process. For a two sided specification, the capability index, Ppk, is calculated as

$$\mathrm{Ppk} = \min\left(\frac{\left(\overline{Y} - LSL\right)}{3 \times S}, \frac{USL - \overline{Y}}{3 \times S}\right) \tag{14}$$

where $\overline{Y}$ represents the process mean, S is the standard deviation, LSL and USL are the lower and upper specification limits, respectively. Using the point estimates in Table 5, the computed Ppk using Eq. (14) is

$$\mathrm{Ppk} = \min\left(\frac{(65.003 - 58.5)}{3 \times 1.179}, \frac{(71.5 - 65.003)}{3 \times 1.179}\right) = \min\left(1.839, 1.837\right) = 1.837 \tag{15}$$

For this example, the two quantities within the parentheses are very close to each other implying that the process is centered within the specification. For this example, assume that a Ppk of 1.5 is acceptable which corresponds to 0.0007 % of the individual values falling outside of the specification limit. The largest mean protein concentration for the post-change process that meets this requirement is computed as follows:

$$\begin{aligned}\mathrm{Ppk} &= \left(\frac{\left(USL - \overline{Y}_{post}\right)}{3 \times S}\right)\\ 1.5 &= \left(\frac{\left(71.5 - \overline{Y}_{post}\right)}{3 \times 1.179}\right)\\ \overline{Y}_{post} &= 66.2\ \mathrm{mg/mL}\end{aligned} \tag{16}$$

Thus, the allowable shift from the given the current pre-change mean is computed as

$$66.2\mathrm{mg/mL} - 65.003\ \mathrm{mg/mL} = 1.197\ \mathrm{mg/mL},$$

and with rounding, the equivalence acceptance criterion is ±1.2 mg/mL. If there were no specification to drive the EAC, a statistical EAC could be considered using the effect size approach. To compute the statistical EAC, Eq. (9) is used. Assuming an effect size of 2 and the standard deviation in Table 5, the statistical EAC equals $\pm 2 \times 1.179 = \pm 2.358$ mg/mL. Note that this is considerably wider than the EAC computed using the specification method. Regardless of the method used to define the EAC, the reasonableness of the EAC should always be evaluated by the subject matter expert.

The next step is to determine the study design or evaluate a proposed study design given a specified number of post-change lots for its power. To evaluate the proposed study design, the power of the study is computed using the EAC, the pre and post-change sample size, and the point estimate for the standard deviation. The following SAS code is used given a true difference of 0 mg/mL with a pre-change sample size of 35 lots and a post-change sample size of three lots.

```
proc power;
twosamplemeans test = equiv_diff
lower = −1.2
upper = 1.2
meandiff = 0
stddev = 1.179
power = .
groupns = (35 3);
run;
```

The power of this test is only 0.069 (6.9 %). This is clearly too low. There are two reasons for the low power. The first reason is that the standard deviation, 1.179, is similar to the EAC, ±1.2. Secondly, the post-change sample size, three lots, is too low given the small allowable shift in means. Since this EAC was computed to control the percent of observations that could fall outside the specification, there is little that can be done other than increasing the post-change sample size. To obtain the desired power of 99 % when the true difference in process means is zero, a post-change sample of 35 lots is required. A sample of 35 is not a reasonable sample size, and the risk of running the study with only three post-change lots is accepted. Should the equivalence test result in an inconclusive result, the burden of proof to demonstrate the pre and post-change processes are comparable is on the manufacturer. It is also important to note that a single inconclusive test result does not "fail" the entire comparability study. The criticality of the parameter relative to the change that is made and the totality of the evidence must be evaluated prior to making a declaration of comparability.

Once the study design has been agreed upon, the post-change data is collected. Although the power was low for this study, only three post-change lots were collected. The resulting equivalence test is conducted using the hypothesis in Eq. (17).

Table 6 Summary of equivalence test for non-profile data

Protein concentration (mg/mL)						
Pre-change mean	Post-change mean	Difference in means	Lower 95 % confidence bound on difference	Upper 95 % confidence bound on difference	EAC	Conclusion
65.0026	65.2910	0.2884	−0.4395	1.0164	±1.2	Statistically equivalent

$$\begin{aligned} H_0 &: |\mu_1 - \mu_2| \geq 1.2 \\ H_a &: |\mu_1 - \mu_2| < 1.2 \end{aligned} \tag{17}$$

To construct the equivalence test, the data are modeled using an independent model described in Eq. (18)

$$\begin{aligned} & Y_{ij} = \mu_i + E_{ij} \\ & i = 1, 2 \ \text{(pre- or post-change process)}\,; \ j = 1, \ldots, n_i \end{aligned} \tag{18}$$

where Y_{ij} is the response for lot j collected from process i, n_i is the number of lots measured from process i $\left(n_1 = 35 \text{ and } n_2 = 3\right)$, μ_1 and μ_2 are the true averages for the pre- and post-changes process respectively, and E_{ij} is the residual error associated with lot j collected from process i. It is assumed that the E_{ij} are sampled from a normal population with mean 0 and variance σ_1^2 for the pre-change process and variance σ_2^2 for the post-change process. The values for the three post-change lots are 65.4631, 64.7484, and 65.6616 mg/mL. Since the lower and upper one-sided 95 % confidence bounds are completely contained within the EAC, the pre and post-change process means are statistically equivalent. The results of the equivalence test are summarized in Table 6.

It is also recommended to plot the results of the equivalence test. Figure 7 consists of two panels. The right-hand panel represents the equivalence test with the EAC. The left-hand panel plots the raw data and the pre and post change means.

3.5 Equivalence Test: Profile Data

The final example describes a test of equivalence with profile data. As in the non-profile example, computation of the EAC, study design and the equivalence test are discussed. The pre-change data are collected for a purity assay over a 3 month time period. Samples are held at the stressed condition of 37°C over 3 months. There are 15 lots in the pre-change data set and samples have been evaluated at 0, 1, 2 and 3 months for each lot. Figure 8 consists of the raw data with the individual predicted slopes fit through each lot where all regression lines are emanating from the average

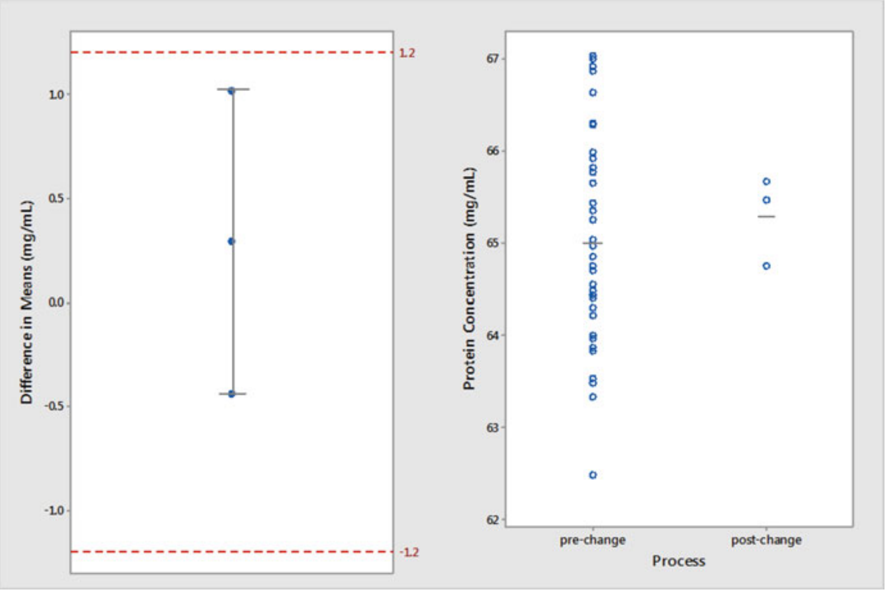


Fig. 7 Graphical summary of the equivalence test for non-profile data with specification for protein concentration (*Left hand panel*: Equivalence plot, *Right hand panel*: Individual value plot with mean)

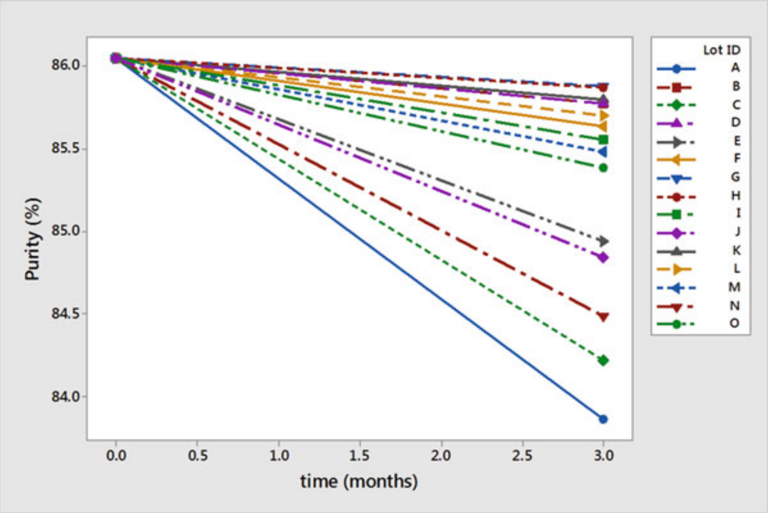


Fig. 8 Pre-change data with lot specific regression lines and common intercept (stressed stability study)

Table 7 Slope estimates (% purity per month) of the pre-change lots

Lot ID	Slope	Lot ID	Slope
A	−0.727	I	−0.165
B	−0.090	J	−0.402
C	−0.610	K	−0.082
D	−0.092	L	−0.115
E	−0.368	M	−0.188
F	−0.137	N	−0.521
G	−0.056	O	−0.220
H	−0.059	–	–

Table 8 SAS output for Example 3

Covariance parameter estimates				
Cov parm	Estimate	Alpha	Lower	Upper
lot	0.4300	0.1	0.2415	1.0236
Residual	0.1991	0.1	0.1448	0.2941

y-intercept of 86.0 % to better visualize the range of slopes for the pre-change data. The slopes range from –0.056 % per month (lot G) to –0.727 % per month (lot A). Table 7 lists the individual slope estimates for each lot.

Since the data are collected from a stressed stability study, a specification does not exist to provide guidance relative to setting the EAC. Therefore, the EAC will be defined using an effect size (Burdick and Sidor 2013). To begin the discussion of selecting an appropriate effect size, Fig. 5 may be used to help the subject matter expert understand the percent overlap between the pre and post-change slopes given a specific effect size. For the example, an effect size of 2 is chosen to compute the EAC. The data are modeled using Eq. (10). The SAS output is listed in Table 8.

The EAC is computed using Eq. (13).

$$\mathrm{EAC} = \mathrm{ES}\sqrt{\frac{S_E^2}{SST}} = 2\sqrt{\frac{0.1991}{5}} = 0.399\ \%\ \text{per month} \tag{19}$$

This EAC needs to be reviewed with the subject matter expert to confirm its reasonableness.

Once the EAC is agreed upon, the comparability study is designed. The pre-change data (15 lots) used to set the EAC will also be used to compute the pre-change slope for the equivalence test. The focus of the study design is the number of post-change lots, the spacing of the time points and replication at specified time points. The proposed study design consists of three post-change lots. The subject matter expert would like to maintain the pre-change time points of 0, 1, 2 and 3 months. A study with no replication and triplicate measurements taken at the 0 and 3 month time is evaluated. Table 9 summarizes power for the two study designs for a true difference in slopes of zero and a one standard deviation shift of the slopes given an EAC of ±0.399 % per month. Since the time points are different

Table 9 Power calculation for design 1 and design 2 for a stressed 3 month study (15 lots pre-change and 3 lots post-change)

Hypothetical shift in slopes	Study design 1: (pre-change: 0, 1, 2, 3) (post-change: 0, 1, 2, 3)	Study design 2: (pre-change: 0, 1, 2, 3) (post-change: 0, 0, 0, 1, 2, 3, 3, 3)
No difference in population slopes	88.7 %	99.6 %

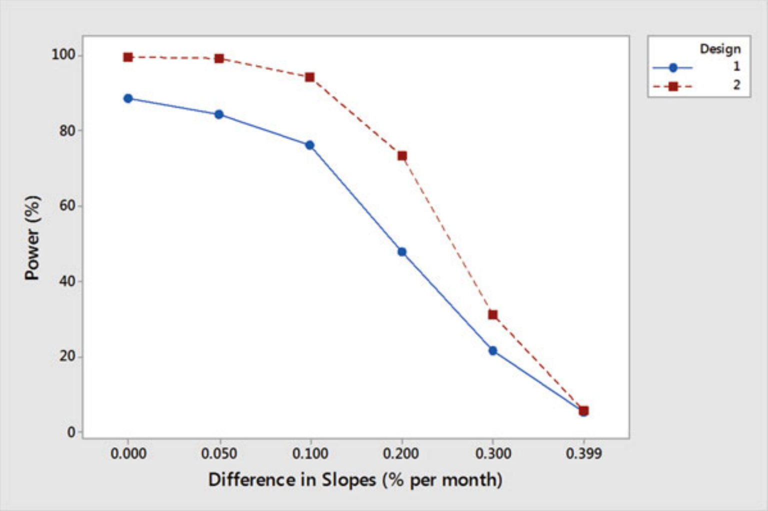


Fig. 9 Power curves for study design 1 and study design 2 for stressed stability study

for each study design, the variance needs to be divided by the appropriate SST to compute the power for the specific study design. The SST for study design 1 and study design 2 is 5 and 14 respectively.

The power curves are plotted for each study design in Fig. 9. In general, it is desired to have a power of at least 99 % when the true difference in slopes is zero. Based on the calculated power, the best study design for the post-change data set is design 2 given the power at a true difference in slopes of zero is at least 99 %. Therefore, study design 2 is chosen over study design 1.

Finally, the equivalence test is performed. This test is performed by computing the two one-sided 95 % confidence bounds (lower and upper) on the difference in slopes. Equivalence is demonstrated if these bounds fall within the range defined by −EAC to +EAC. The model that is fit to perform the equivalence for profile data is

$$Y_{ijkm} = \mu_i + L_{j(i)} + \gamma_i \times t_{ijkm} + E_{ijkm} \tag{20}$$

where Y_{ijkm} is the m^{th} replicate of the k^{th} time point for lot j sampled from process i; μ_i is the mean of process i; $L_{j(i)}$ is a normal random variable that allows the response at the initial time point to vary across lots with mean 0 and variance σ_L^2; γ_i is the slope of process i; t_{ijkm} is the time point for the m^{th} replicate of the k^{th} time point for lot j sampled from process i; E_{ijkm} is a random normal error term with mean 0 and variance σ_E^2; and $L_{j(i)}$ and E_{ijkm} are jointly independent. Note that Eq. (20) may be used to model stressed stability data that have different time points for the different lots. The equivalence test defines the set of hypotheses as

$$\begin{aligned} H_0 &: \left|\gamma_{pre} - \gamma_{post}\right| \geq 0.399 \\ H_a &: -0.399 < \gamma_{pre} - \gamma_{post} < 0.399. \end{aligned} \tag{21}$$

The data are listed in the Appendix.

Table 10 provides a tabular summary of the equivalence test. In Fig. 10, the left-hand plot is the equivalence plot and the right-hand plot consists of the raw data and

Table 10 Tabular summary of equivalence test results for stressed stability study

Purity (% per month)						
Pre-change slope	Post-change slope	Difference in slopes	Lower 95 % confidence bound on difference	Upper 95 % confidence bound on difference	EAC	Conclusion
−0.2555	−0.4651	−0.2096	−0.3432	−0.0761	±0.399	Statistically equivalent

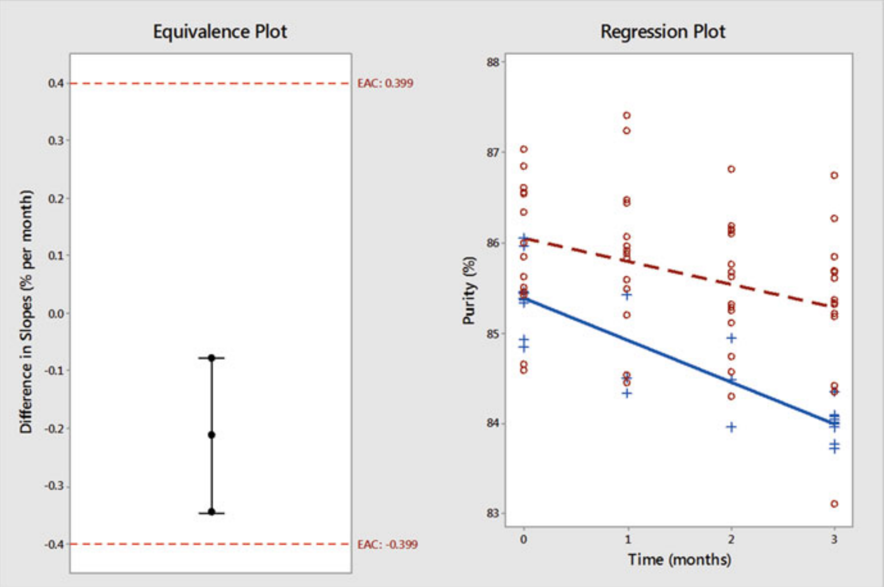


Fig. 10 Graphical summary of equivalence results for Example 5 (Post-change: *solid line* with '+' symbols and Pre-change: *dashed line* with open circles)

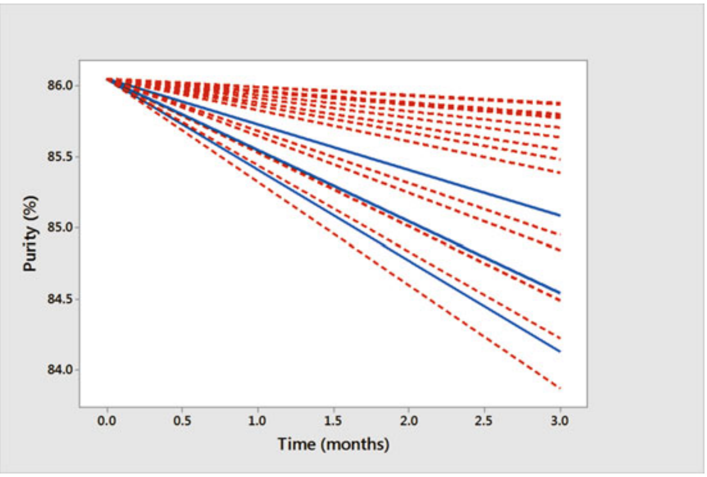


Fig. 11 Individual regression lines by process (*blue solid line*: post-change process and *red dashed line*: pre-change lots)

the slopes for each process. Since the lower and upper 95 % one-sided confidence bounds fall within the EAC, the two processes are statistically equivalent.

Another helpful plot to consider is to graph the slopes of all pre and post-change lots with a common y-intercept (Fig. 11). This type of plot aids in the visualization of the slope overlap.

As with all of the examples presented, more complex models may be used to estimate the variance components. For profile data, not only is the data structure important to consider, but the reaction kinetics is critical in selecting the most appropriate model.

4 Conclusion

Regulatory bodies recognize the importance of continuous improvement in a manufacturing setting and acknowledge that every change does not merit a clinical or non-clinical study. This flexibility in demonstrating comparability is outlined in ICH Q5E. However, specific guidance on how to compare the pre and post-change process data is not made when evaluating analytical comparability. This paper provides statistical and non-statistical recommendations on setting acceptance criteria based on the desired definition of comparability. If comparability is more logically defined in terms of estimable process and product parameters (e.g. means and slopes), the recommended approach is an equivalence test. If the comparison of parameters is not of interest, multiple options are available (statistical intervals or other visualization techniques). However, these options are not statistical tests.

Examples of statistical intervals presented in this paper are the use of tolerance or prediction intervals. If computing intervals, the interval must be within the specification limit to be useful. Regardless of the methodology chosen, the statistician needs to be aware of the data structure and the variance components, and the subject matter expert must provide their expertise in assessing the reasonableness of the criteria. Finally, regardless of the proposed methodology, an analytical comparability study must include a study design, and the corresponding acceptance criteria must be defined prior to collecting the post-change data.

Disclaimers The thoughts and opinions presented in this article represent the author's positions.

Software Software used for the computations in this chapter are Minitab v17.0, SAS University Edition v3.2 and MS Excel.

A.1 Raw Data

Example 1: tolerance interval

Lot ID	Clinical scale data	Lot ID	Pre-change commercial scale data
A	33.2111	F	54.0648
B	37.5348	G	59.7112
C	36.1102	H	55.5946
D	35.0890	I	59.5768
E	34.9719	J	52.8764

Example 2: non-profile equivalence

Lot ID	Value	Lot ID	Value	Lot ID	Value	Lot ID	Value	Lot ID	Value
1	64.6901	8	63.8596	15	63.4737	22	65.2484	29	65.7704
2	66.2940	9	64.4832	16	64.3895	23	66.6342	30	63.8568
3	65.8114	10	66.9188	17	63.5307	24	65.3451	31	63.9972
4	65.6446	11	64.8376	18	65.0348	25	66.8672	32	65.9727
5	63.9546	12	64.9620	19	66.9927	26	63.8307	33	64.5528
6	66.2753	13	64.7369	20	67.0247	27	65.9205	34	64.2946
7	64.4325	14	63.3229	21	62.4785	28	64.2135	35	65.4374

Example 3: profile equivalence

Lot ID	Time point	Value	Process	Lot ID	Time point	Value	Process	Lot ID	Time point	Value	Process
A	0	85.450	PRE	H	0	85.420	PRE	O	0	86.340	PRE
A	1	84.540	PRE	H	1	85.200	PRE	O	1	86.430	PRE
A	2	84.290	PRE	H	2	85.240	PRE	O	2	86.180	PRE
A	3	83.110	PRE	H	3	85.210	PRE	O	3	85.690	PRE
B	0	85.500	PRE	I	0	84.580	PRE	P	0	86.050	POST
B	1	85.820	PRE	I	1	85.910	PRE	P	0	85.380	POST
B	2	85.310	PRE	I	2	84.740	PRE	P	0	85.970	POST
B	3	85.369	PRE	I	3	84.420	PRE	P	1	84.500	POST
C	0	86.340	PRE	J	0	86.610	PRE	P	2	84.940	POST
C	1	86.070	PRE	J	1	87.410	PRE	P	3	84.080	POST
C	2	85.760	PRE	J	2	85.670	PRE	P	3	83.770	POST
C	3	84.410	PRE	J	3	85.850	PRE	P	3	84.100	POST
D	0	86.000	PRE	K	0	84.650	PRE	Q	0	85.450	POST
D	1	85.870	PRE	K	1	84.450	PRE	Q	0	85.370	POST
D	2	86.150	PRE	K	2	84.560	PRE	Q	0	85.330	POST
D	3	85.600	PRE	K	3	84.340	PRE	Q	1	85.420	POST
E	0	86.840	PRE	L	0	86.540	PRE	Q	2	84.480	POST
E	1	85.480	PRE	L	1	86.440	PRE	Q	3	83.720	POST
E	2	85.280	PRE	L	2	86.100	PRE	Q	3	84.050	POST
E	3	85.680	PRE	L	3	86.270	PRE	Q	3	83.990	POST
F	0	85.620	PRE	M	0	85.850	PRE	R	0	85.430	POST
F	1	85.590	PRE	M	1	85.970	PRE	R	0	84.840	POST
F	2	85.120	PRE	M	2	85.620	PRE	R	0	84.930	POST
F	3	85.320	PRE	M	3	85.340	PRE	R	1	84.330	POST
G	0	86.560	PRE	N	0	87.030	PRE	R	2	83.950	POST
G	1	87.240	PRE	N	1	86.470	PRE	R	3	84.350	POST
G	2	86.140	PRE	N	2	86.810	PRE	R	3	83.950	POST
G	3	86.740	PRE	N	3	85.180	PRE	R	3	84.010	POST

References

FDA Guidance Concerning Demonstration of Comparability of Human Biological Products, Including Therapeutic Biotechnology-derived Products (April 1996).

ICH Guidance for Industry: Q5E Comparability of Biotechnology/Biological Product Subject to Changes in Their Manufacturing Process (June 2005).

Chatfield, M.J., Borman, P.J.; Damjanov, I.; “Evaluating Change During Pharmaceutical Product Development and Manufacture – Comparability and Equivalence”; *Quality and Reliability Engineering International*; 27 (2011), pp 629–640.

Howe, W.G.; Two-Sided Limits for Normal Populations – Some Improvements; *Journal of the American Statistical Association*, 64 (1969), pp. 610–620.

Hahn, G.J. and Meeker, W.Q.; Statistical Intervals A Guide for Practioners; John Wiley & Sons, Inc.; (1991).

Limentami, G.B, et al; "Beyond the t-Test: Statistical Equivalence Testing", *Analytical Chemistry*; June (2005), pp 221–226.

Schuirmann, D.J. A comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability; *Journal of Pharmoacokinetics*, Vol. 15 (1987), No. 6, pp 657–680.

Chatfield, M.J. and Borman, P.J.; "Acceptance Criteria for Method Equivalency Assessments", Analytical Chemistry; 81 (2009), pp 9841–9848.

Hauk, W.W., Abernethy, D.R., and Williams, R.L. (2008); Metrological approaches to setting acceptance criteria: Unacceptable and unusual characteristics; *Journal of Pharmaceutical and Biomedical Analysis*, 48, pp 1042–1045.

Lei, L. and Olson, K.; "Evaluating Statistical Methods to Establish Clinical Similarity of Two Biologics", *Journal of Biopharmaceutical Statistics*; 20: 1 (2010), pp 62–74.

Burdick, R.K. and Sidor, L.; "Establishment of an Equivalence Acceptance Criterion for Accelerated Stability Studies"; Journal of Biopharmaceutical Statistics; 23: 4 (2013), pp 730–743.

Statistical Applications for Biosimilar Product Development

Richard Montes, Bryan Bernat, and Catherine Srebalus-Barnes

Abstract Regulatory approval of biosimilar products requires demonstration of analytical similarity of functional and structural attributes between the proposed biosimilar product and on-market reference product. The statistical framework for how to evaluate the analytical similarity data has recently been published and a U.S. regulatory guidance is expected soon. This paper illustrates the challenges and issues encountered by Hospira (a Pfizer company) in implementing this newly described statistical framework to support the analytical similarity assessments for biosimilar products. A simulation approach using multilevel (hierarchical) linear regression is also proposed to statistically derive shelf-life specification limits. The approach may be applicable when there is larger volume of data that can be generated as part of the analytical similarity assessment. The performance of the simulation approach is compared when there is a limited vs. sufficiently large sample size and when the quality attribute of interest has a low vs. high analytical variability. The proposed simulation approach to calculate shelf-life specification limits is also benchmarked against a commonly utilized approach in industry based on a fixed effect Analysis of Covariance (ANCOVA) model.

Keywords Biosimilars • Equivalence testing • Shelf-life specification limits • ANCOVA • Hierarchical linear regression

1 Introduction

The approval of the first biosimilar (FDA 2015) by the FDA in March 2015 is an historic event for the U.S. healthcare system. There are several biosimilar product applications currently under FDA review while the number of biosimilar candidates being developed by sponsor companies continues to increase. The 2015 FDA draft biosimilars guidance document (Guidance for Industry 2015) recommends that

R. Montes (✉) • B. Bernat • C. Srebalus-Barnes
Hospira, a Pfizer company, 275 N. Field Drive, Bldg H3-3N, Lake Forest, IL 60045, USA
e-mail: richard.montes@pfizer.com; bryan.bernat@pfizer.com;
catherine.srebalusbarnes@pfizer.com

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_20

sponsors use a stepwise approach to establish the totality of the evidence that supports a demonstration of biosimilarity. The stepwise approach starts with extensive structural and functional characterization of both the proposed biosimilar product and the reference product to demonstrate their analytical similarity. If any residual uncertainly about biosimilarity remains after this initial characterization, additional pre-clinical and clinical studies are performed. The focus of this paper is on the analytical similarity assessment from the structural and functional characterization. Development of appropriate statistical methods to support the analytical similarity assessment that provide statistical rigor yet accommodate the challenges and issues encountered in biosimilar product development is the primary challenge for the statisticians supporting both regulatory agencies and sponsors companies.

In addition to the demonstration of analytical similarity, the Chemistry, Manufacturing and Control (CMC) requirements for new biologic product licensing applications also apply for biosimilar product applications. Requirements such as stability analyses for shelf-life estimation and specification setting rely heavily on application of statistical methodologies. As a consequence of the early characterization work requisite for the analytical similarity assessment, the number of biosimilar lots studied for characterization and stability studies can often be considerably larger than the minimum of three stability lots required for a new biologic product. The availability of a larger biosimilar data set enables the statistician to apply statistical methodologies with improved precision and reliability.

This paper covers two main topics of statistical applications for biosimilar product development. First, the challenges and issues encountered in the implementation of the FDA approach for evaluation of analytical biosimilarity data will be illustrated. Second, shelf-life specification limits set using hierarchical (multilevel) linear regression modeling simulation will be compared when there is a limited vs. large number of stability lots and when the quality attribute has low vs. high analytical variability.

2 Challenges and Issues in Implementing Tier 1 Equivalence Testing

The first focus of this article is to describe the challenges and issues encountered in implementing the FDA analytical similarity assessment framework described in Tsong et al. (2015). The framework starts with ranking the critical quality attributes into three statistical analysis tiers based on the potential impact of the attribute on product quality and clinical outcomes. Different statistical approaches with varying degrees of rigor commensurate to the criticality of the attribute are applied for attributes assigned to the different statistical tiers. The different statistical approaches recommended in decreasing order of statistical rigor are equivalence testing for Tier 1, quality range assessment for Tier 2, and descriptive raw data and graphical comparison for Tier 3.

The Tier 1 equivalence testing uses the two one-sided tests (TOST) on the mean difference of test and reference products ($\mu_T - \mu_R$) described in Westlake (1981) and Schuirmann (1987). TOST tests the two sets of one-sided hypotheses decomposed from the null H_0 (inequivalence) and alternate H_1 (equivalence) interval hypotheses (Eq. 1). Note that the formulated inequivalence and equivalence hypotheses may be interpreted as "analytically dissimilar" and "analytically similar", respectively, in the context of biosimilarity assessment. Operationally, the TOST procedure is the same as declaring equivalence if the $100\,(1 - 2\alpha)\,\%$ two-sided confidence interval for ($\mu_T - \mu_R$) is completely contained in the equivalence margins $[-\delta, \delta]$. The equivalence margins $[-\delta, \delta]$ ideally are established prospectively by studying the maximum mean difference between the products that do not have a clinically significant impact on safety and efficacy. In practice, a clear linkage between analytical differences and clinical impact is typically not well-defined. Tsong et al. (2015) used a power-based calculation to set the equivalence margins. For example, if the true mean difference between test and reference products is 1/8 of the standard deviation of the reference product σ_R and the numbers of lots are 10 each, setting $\delta = 1.5\sigma_R$ results in 87 % power of the TOST. This method of establishing equivalence margins does not take into consideration the shift in the means of the two products that has no clinical impact. There is expectation that this clinically insignificant shift in product means, although unknown, realistically exists and is likely larger than the $\frac{1}{8}\sigma_R$ shift provided in Tsong et al. (2015). The Tier 1 equivalence margin setting approach is therefore considered a conservative approach.

$$\begin{array}{lll} H_0\ (inequivalence): & H_{01}:\ \mu_T - \mu_R \le \delta_L; & H_{02}:\ \mu_T - \mu_R \ge \delta_U \\ H_1\ (equivalence): & H_{11}:\ \mu_T - \mu_R > \delta_L; & H_{12}: \mu_T - \mu_R < \delta_U \end{array} \tag{1}$$

The Tier 2 assessment evaluates whether a sufficiently high proportion (e.g., 90 %) of the biosimilar lots fall within a quality range. The quality range is defined from the mean $\left(\overline{\overline{X}}_{reference}\right)$ and the observed standard deviation of the lot means $\left(s_{\overline{X}_{reference}}\right)$ of reference product, formulated analogously as control chart limits, $\left[\overline{\overline{X}}_{reference} \pm c * s_{\overline{X}_{reference}}\right]$. The c multiplier is justified for each Tier 2 attribute analyzed.

The challenges and issues described in the following sub-sections are geared more specifically towards Tier 1 analysis.

2.1 Reference Product Variability to Define Equivalence Margin Is Estimated from the Sample

The first challenge encountered in implementing the Tier 1 analysis is the estimation of true variability of the reference product σ_R. In the power calculation described above, it is assumed that σ_R is known. In practice, it is unknown and actually estimated as $\hat{\sigma}_R$ (standard deviation of the reference product sample) from the same sample used to perform the equivalence test. Burdick and Ramírez (2015) and Zhang and Wu (2015) showed through simulation that such practice reduces the purported power from 87 to 80 % and inflates Type I error from 5 to 6 % of TOST for the sample size of 10 lots each product and using $\delta = 1.5\hat{\sigma}_R$. This decreases the probability that a biosimilar product truly analytically equivalent to a reference product can be concluded as such. It also increases the risk of declaring an analytically dissimilar biosimilar product as analytically similar to a reference product using the TOST analysis.

2.2 Tier 1 Attribute Assignments

Tier 1 attributes are designated as those most relevant to the therapeutic mechanism of action for the biosimilar and reference products. The selection of Tier 1 attributes can be challenging in cases where an extensive array of functional assays is used to comprehensively probe one or more possible mechanisms of action. The multiple functional attributes evaluated for the analytical biosimilarity assessment may be overlapping or redundant. For example, *in vivo* biopotency, *in vitro* biopotency, and receptor binding attributes all measure the ability of the product to activate various pathways in the therapeutic mechanism of action. A multivariate analysis of these three attributes show strong pairwise correlation with each other (Fig. 1) which confirms that they are indeed overlapping attributes. Among these functional assays, *in vivo* biopotency is considered the most relevant to the mechanism of action for the specific biosimilar product evaluated because it includes an assessment of *in vivo* clearance pathways that may impact availability of the therapeutic protein at the receptor to initiate the relevant biological pathways. *In vivo* biopotency is therefore assigned as a Tier 1 attribute to be analyzed using TOST while the redundant functional attributes are assigned as Tier 2 to be evaluated using the quality range assessment. Limiting the attributes classified as Tier 1 to a small number focuses statistical assessment to the most important attributes and reduces potential for confounding results from multiple Tier 1 analyses.

2.3 Age Matching of Reference and Biosimilar Products

In performing analytical similarity assessment, it is ideal to match the age of reference and biosimilar products so as to make the comparison as fair as possible. Procuring the reference product for the structural and functional characterization is a logistical challenge. Since there is a lag between the dates a reference lot is manufactured and when it becomes commercially available for purchase, it is generally not feasible to match the age of the biosimilar product and the reference product at the time of testing. Further, the window of opportunity for testing a

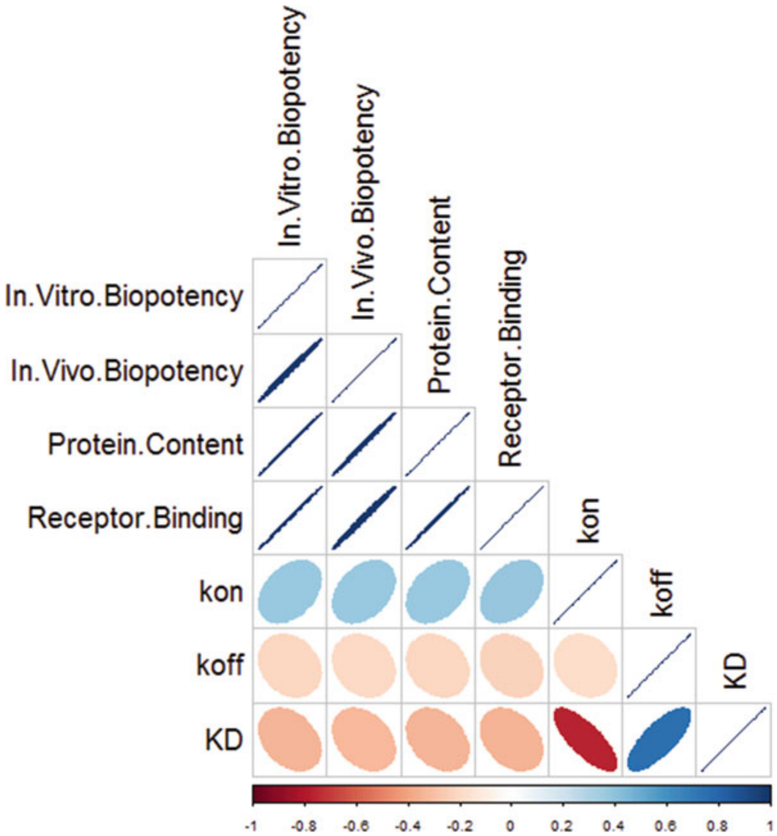


Fig. 1 Pairwise correlation plots between various functional assays

procured reference product is limited by whatever remaining shelf-life it has from the date of manufacture, a date which is usually unknown and only approximated. Additionally, reference and biosimilar products need to be characterized over time to infer any temporal trend which further complicates the age-matching requirement by the FDA.

To illustrate these challenges, consider the simulated lot measurements over time of a particular attribute for biosimilar and reference products in Fig. 2. If product procurement and testing resources are not limiting, numerous lots may be manufactured or purchased and tested at multiple different time points across the product shelf-life (Fig. 2, left) providing sufficient data to confidently conclude that there is no systematic trend over time. Equivalence testing with age-matching between the biosimilar and reference products can be performed. If only a few biosimilar (or reference) lots can be manufactured (or procured) and testing is limited (Fig. 2, right), the data is not amenable to meaningful statistical inference on temporal trends and age-matching may not be possible. A pragmatic approach

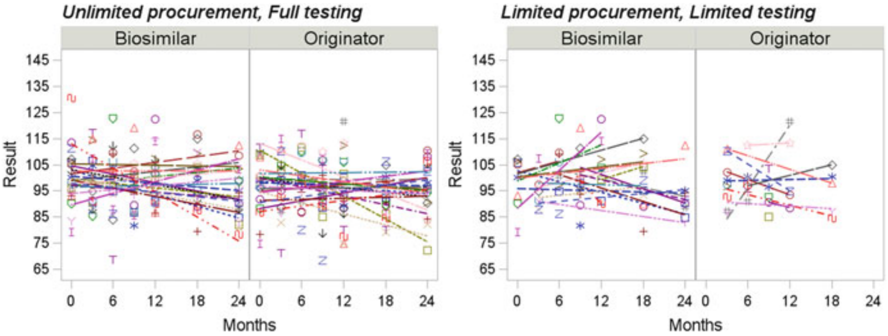


Fig. 2 Simulated lot measurements of a quality attribute over time for biosimilar and reference products with fitted regression lines by lot when there is Unlimited reference product lots procured and fully tested (*left*) and when there is limited reference product lots procured and limited testing for both products (*right*)

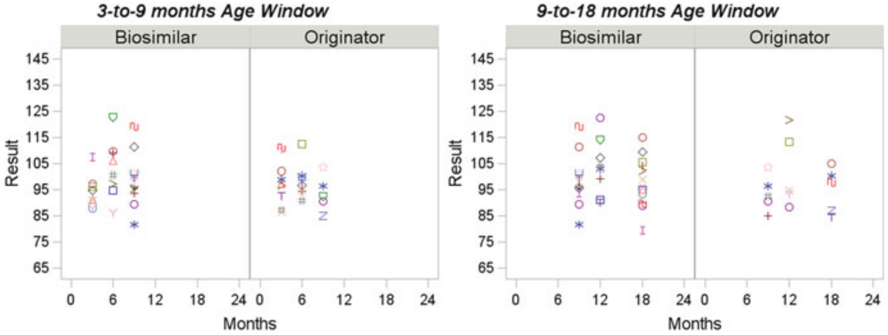


Fig. 3 Simulated data from Fig. 2 (*right*) sub-grouped into 3–9 months age window (*left*) and 9–18 months age window (*right*)

that may be adopted to address these issues is to group the entire dataset into "age windows" aligned with the age range of the products at the time of testing. Age windows are arbitrarily defined as product age ranges that are narrow enough such that any temporal trend, if any, can be considered negligible relative the overall temporal trend in the entire data set. The age windows are chosen so that there will be sufficient age window-matched data points for the two products. In this simulated example, age windows of 3–9 months (Fig. 3, left) and of 9–18 months (Fig. 3, right) are defined. Because the attribute is considered time-invariant over these windows, the lot measurements can serve as pseudo lot replicates which are averaged and reported for the lot. The lot means are used for the TOST analyses and the results are shown in Fig. 4 for the 3–9 month age window (left) and 9–18 month age window (right). Equivalence is concluded for both age windows. The age window approach applied meets the FDA recommendation of age-matching for the equivalence testing while providing a pragmatic solution for having limited data due to procurement and testing resources issues.

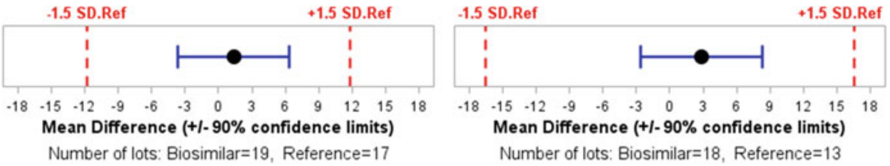


Fig. 4 Equivalence testing results for 3–9 months age window (*left*) and 9–18 months age window (*right*)

2.4 *The Number of Reference and Biosimilar Lots Should Be Balanced*

It is ideal to have an approximately balanced number of reference and biosimilar product lots used in the analytical similarity assessment. This is because the probability of passing equivalence testing can be unduly increased by simply increasing the sample size of one of the product types as can be shown through simulation (results not shown). In reality, procurement and testing constraints as well as age-matching dictate the final number of lots available for the similarity assessment. In the simulated example for the 9–18 age window ultimately used for TOST (Fig. 4, right), there are 18 biosimilar lots vs. 13 reference lots. The direction of inequality can go either way in practice. If the reference lots outnumber the biosimilar lots, "splitting practice" is described in Chow (2014) wherein a portion of reference lots is set aside to estimate σ_R to define δ and the remainder is used for the TOST analyses. Burdick and Ramírez (2015) showed through simulation that such practice decreases the statistical power and inflates the Type I error of TOST, which compounds the problem described when $\hat{\sigma}_R$ estimated from the reference sample is used to define the equivalence margins (see Sect. 2.1). If a perfectly balanced number of lots is enforced as recommended by the FDA, the challenge is to set an objective algorithm on how to balance the number of lots. Taking the 18 biosimilar lots 13 at a time to balance the number of reference lots in Fig. 4 (right) leads to 8568 possible TOST data sets! Not using all the generated data for the sake of balancing the number of lots goes counter to the statistical thinking that the more the samples drawn from a population, the better the inferences on its mean, spread, and shape that all factor into the TOST analysis.

The challenges and issues described above which are statistical and/or logistical in nature are just some encountered in implementing the analytical similarity assessment statistical framework recommended by FDA. Currently, there are continuing statistical research and dialogue between sponsors and agency to address issues such as impact of defining equivalence margins from sample-estimated reference product variability, alternative to the mean difference between products used as test criterion in TOST, and how to handle unbalanced data sets.

3 Statistical Applications for Chemistry, Manufacturing and Control (CMC) Requirements of Biosimilars

In addition to the analytical biosimilarity assessment, the standard CMC requirements for new biologic product licensing applications are also needed for biosimilar product applications. Specification limits are one component of the overall control strategy to ensure product quality and consistency. Specifications are numerical limits to which a product must conform at the time of manufacture (release specification limits) and throughout the product shelf-life (shelf-life specification limits) to be considered acceptable. Specification limits are justified based on overall review of the development and manufacturing experience for the product as described in guidance documents Q6A (2000) for small molecule products and Q6B (1999) for biological products.

3.1 Relationship Between Setting Specification Limits and Establishing the Product Shelf-Life

The specification limits are directly linked to the proposed product shelf-life. The product shelf-life is the time period during which a drug product is expected to remain within the approved shelf-life specifications as defined in Q1A(R2) (2003). Shelf-life is set by statistical analysis of stability data (i.e., time-profiles for quality attributes of lots sampled from the product population) as described in Q1E (2004). The lots are modeled as a fixed effect with time as covariate using analysis of covariance (ANCOVA) with specified rules for testing the poolability of slopes and intercepts as detailed in Q1E. The time at which a 95 % confidence interval of the predicted mean and the proposed shelf-life specification limits intersects is set as the maximum shelf-life that can be supported for the product.

There are no set statistical methods prescribed in guidance documents Q6A and Q6B to derive specification limits. Because the shelf-life determination is closely associated with the shelf-life specifications, the statistical methods in the former as detailed in Q1E are sometimes extended to the latter by some industry practitioners. For example, a provisional product shelf-life would be set first using non-statistical considerations (e.g., business plans, prior knowledge on related products, etc.). The statistical exercise at hand then is to predict the future values at the provisional shelf-life using the calculations from Q1E to inform what shelf-life specification limits to set. The problem with this approach is that the ANCOVA method in Q1E models lot as fixed effect and therefore inferences can only be applied for the particular lots monitored in the stability studies. It is assumed that the monitored lots are representative of the population. However, given that the minimum number of stability lots to perform the shelf-life expiry analysis for a new biologic is three (which may be all that is available at time of drug filing), it is difficult to check the assumption that these lots are truly representative of the population. In addition, the inferences for shelf-life calculations in Q1E are based on the predicted mean values whereas lot release and stability measurements assessed against specification limits using the individual results.

3.2 Need for Statistical Methodology for Specification Setting That Accounts for Random Lot Effect

Since there are no set methods prescribed in guidance documents Q6A and Q6B for setting statistically-derived shelf-life specification limits based on available batch data, sponsors must justify the approach they employ. Some published literature for specific methodologies or general considerations for setting specification limits statistically are briefly surveyed. Allen et al. (1991) developed a practical method to set release specification limits but it is based on the fixed effect ANCOVA modeling outlined in Q1E. Murphy and Hofer (2002) accounted for the random lot effect in calculating the changeover shelf-life expiry for setting release and shelf-life specification limits effect but only focused on random slopes and not random intercepts. Schofield et al. (2008) discussed various issues to consider in a rational approach for setting specification. More recently, Dong et al. (2014) prescribed confidence limits of percentile as a more appropriate alternative to β-content tolerance intervals to set specification limits when there are limited data. However, the approach only applies to univariate data (e.g., lot release data) and not to time-profile data (i.e., stability studies).

In this paper, it is assumed that statistically derived specification limits will be calculated from available release and stability data. It is further assumed that the time profile in the stability data can be described by zero-order kinetics so linear regression is applicable. Given this data structure, an applicable choice for analysis is the multilevel (hierarchical) linear regression (Gelman and Hill 2007). Stability data can be classified as hierarchical because the time measurements are grouped under the particular lots studied. We are interested in inferring about the population from which the lots were randomly sampled from. The random lot effect manifests in both the intercepts and the slopes so the multilevel (hierarchical) linear regression method is also called random coefficient model (RCM). The RCM may not be very useful if there are only limited stability lots as is the case with new biological entity with only three lots. However for biosimilar product development, the available data can be considerably larger than the minimum of three stability lots required for a new biologic. This is because analytical biosimilarity assessment requires characterizing numerous lots of biosimilar product early on to demonstrate equivalence with reference product. The larger body of data obtained in biosimilar product development increases the applicability of RCM for describing the time profiles of biosimilar lots.

3.3 Simulation Using Random Coefficient Model for Shelf-Life Specification Setting

This section uses simulation to illustrate the derived benefits as it relates to setting shelf-life specification limits when there are more time-profile data generated as part of biosimilar characterization vs. when there is only smaller data for a new biologic entity. The sample sizes simulated for this comparison are three stability lots to represent what is encountered in new biologic product development and 20 stability lots to represent biosimilar product development.

Consider a time-dependent quality attribute described by $Y_{ij} \sim N\left(\alpha_i + \beta_i t_j,\ \sigma_Y^2\right)$, where Y_{ij} is a measurement of a product quality attribute from the ith lot ($i = 1, \ldots, I\ lots$) at the j$^{\text{th}}$ t time point ($j = 0,\ 3,\ 6,\ 9,\ 12,\ 18,\ and\ 24$) months, α_i is the intercept of the i$^{\text{th}}$ lot as drawn from normal distribution with fixed mean intercept μ_α and standard deviation σ_α, β_i is the slope of the ith lot as drawn from normal distribution with fixed mean slope μ_β and standard deviation σ_β, and σ_Y is the standard deviation of the error for measuring Y_{ij}. Stability lots of sample sizes 3 or 20 lots are hypothetically generated by drawing random samples from the Y_{ij} normal population with the following parameter values: $\mu_\alpha = 100$, $\sigma_\alpha = 5$, $\mu_\beta = -0.1$, $\sigma_\beta = 0.01$, and σ_Y either at 3 or 12. The scenario $\sigma_Y = 3$ represents a quality attribute whose measurement method has a low analytical variability while $\sigma_Y = 12$ is for one with a high analytical variability. The hypothetical stability data are fitted with Eq. (2) where ε_Y is residual error with distribution $N(0, \sigma_Y^2)$ to obtain $\hat{\mu}_\alpha$, $\hat{\sigma}_\alpha$, $\hat{\mu}_\beta$, $\hat{\sigma}_\beta$, and $\hat{\sigma}_Y$ point estimates of the RCM.

$$Y_{ij} = \alpha_i + \beta_i t_j + \varepsilon_Y \tag{2}$$

In product commercialization, data at release ($t = 0\ months$) would also have been generated along with stability data. Release data typically has a larger lot sample size than the stability data so release data, R_i, of sample size 30 is simulated from distribution $\sim N\left(\mu_R, \sigma_R^2\right)$ where μ_R is the fixed mean of release (also equal to μ_α) and σ_R is the standard deviation of release (equal to $\sqrt{{\sigma_\alpha}^2 + {\sigma_Y}^2}$).

Values of 10,000 future measurements of the attributes evaluated at the proposed end of shelf-life $t_{expiry} = 24\ months$ are simulated by drawing normal random variates from the parameter estimates as formulated in Eq. (3) where $a \sim N\left(\hat{\mu}_\alpha, \hat{\sigma}_\alpha^2\right)$, $b \sim N\left(\hat{\mu}_\beta, \hat{\sigma}_\beta^2\right)$ and $e \sim N\left(0, \hat{\sigma}_Y^2\right)$. The simulated intercept (i.e., a in Eq. 3) may be more accurate and more precise if based on the parameter estimates from a univariate analysis of the release data than from the hierarchical linear regression modeling of the stability data. This is because release data likely has a larger sample size than stability data which can be as few three lots. The a intercept term can then be alternatively simulated from $r \sim N\left(\hat{\mu}_R, \hat{\sigma}_R^2\right)$. The limitation of this approach is that the uncertainty in estimating the parameters is not accounted for.

$$Y_{future\ \ values\ \ at\ \ t=expiry} = a + b * t_{expiry} + e \tag{3}$$

From the 10,000 simulated future values evaluated at shelf-life expiry, the 0.997 central quantile limits (i.e., 0.0015 lower and 0.9985 upper quantiles) of the distribution are assigned as the calculated shelf-life specification limits. The 0.997 quantile corresponds to the intended coverage of the population of future values. This entire sequence of generating hypothetical release and stability data sets, estimation of parameter estimates, simulating 10,000 future shelf-life expiry values based on point estimates of RCM, and outputting the 0.997 quantiles is repeated 100 times. The effects of having either 3 or 12 sampled stability lots for a quality attribute with either low ($\sigma_Y = 3$) or high ($\sigma_Y = 12$) analytical variability are evaluated using two metrics. The first is the *mean of the parameter estimates* from the 100 repetitions; the closer the mean estimates are to the true population parameters, the better. The second is the *confidence coefficient* which is the proportion of the 100 repetitions that the parameter estimate correctly contains the true population parameter; the closer the confidence coefficient to the nominal target of 0.95, the better. All analyses are performed using R 3.2.2 statistical software with library 'arm' package and 'lmer' function.

The $\hat{\mu}_\alpha$, $\hat{\sigma}_\alpha$, $\hat{\mu}_\beta$, $\hat{\sigma}_\beta$, and $\hat{\sigma}_Y$ estimates and the resulting calculated shelf-life specification limits plotted for the 100 repetitions are presented in Figs. 5, 6, 7, and 8 for the four $\left(n_{stability} = 3 \ and \ 20 \ lots\right)$ by ($\sigma_Y = 3 \ and \ 12$) permutations. The true population parameters and true shelf-life specification limits (Eq. 4) are overlayed as red reference lines in the plots. The mean estimates and confidence coefficients are annotated below each plot. The true parameter values, mean estimates, and confidence coefficients are also summarized in Tables 1 and 2 for $\sigma_Y = 3$ and

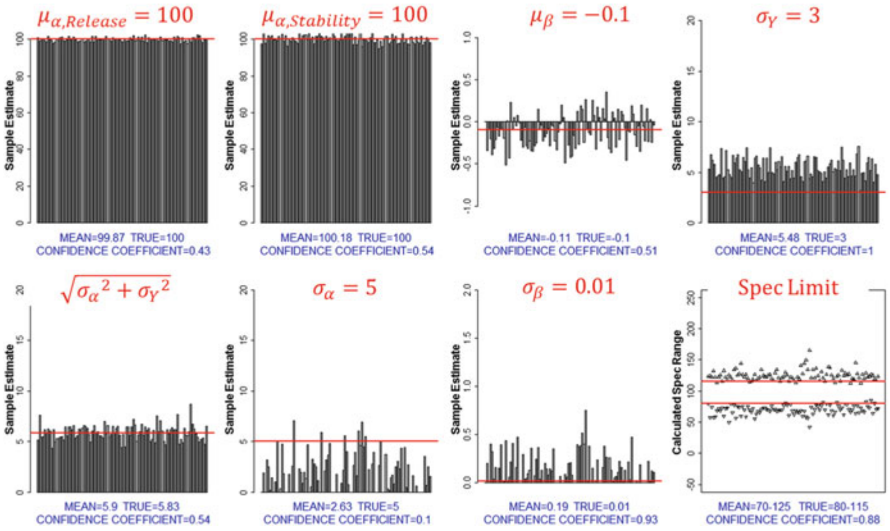


Fig. 5 Bar graphs of parameter estimates and resulting shelf-life specification limits from 10,000 simulations for $n_{stability} = 3$ and $\sigma_Y = 3$. *Note:* x-axis is the index for 100 repeats of simulation; *red reference lines* are the TRUE parameter values

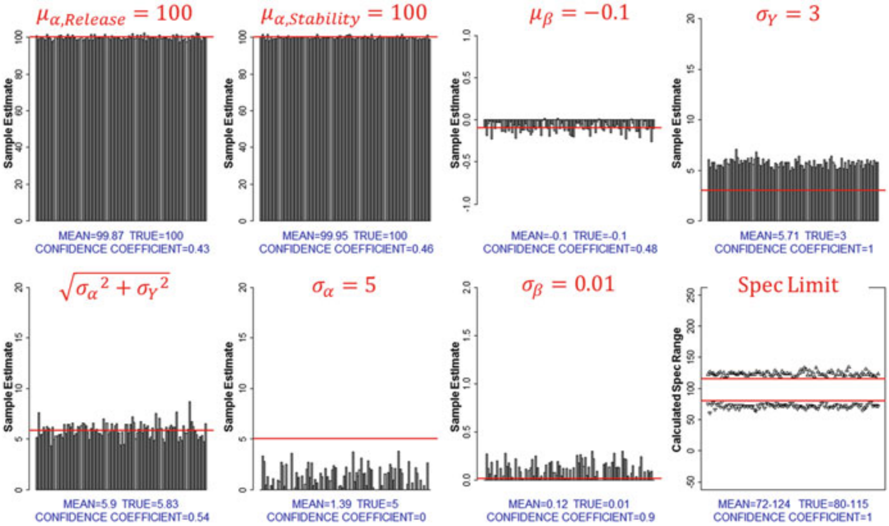


Fig. 6 Bar graphs of parameter estimates and resulting shelf-life specification limits from 10,000 simulations for $n_{stability} = 20$ and $\sigma_Y = 3$. *Note:* x-axis is the index for 100 repeats of simulation; *red reference lines* are the TRUE parameter values)

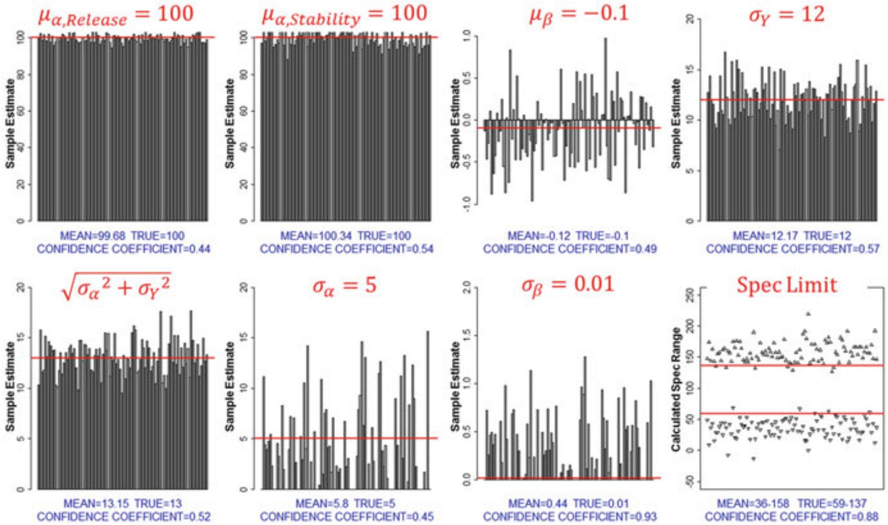


Fig. 7 Bar graphs of parameter estimates and resulting shelf-life specification limits from 10,000 simulations for $n_{stability} = 3$ and $\sigma_Y = 12$. *Note*: x-axis is the index for 100 repeats of simulation; *red reference lines* are the TRUE parameter values

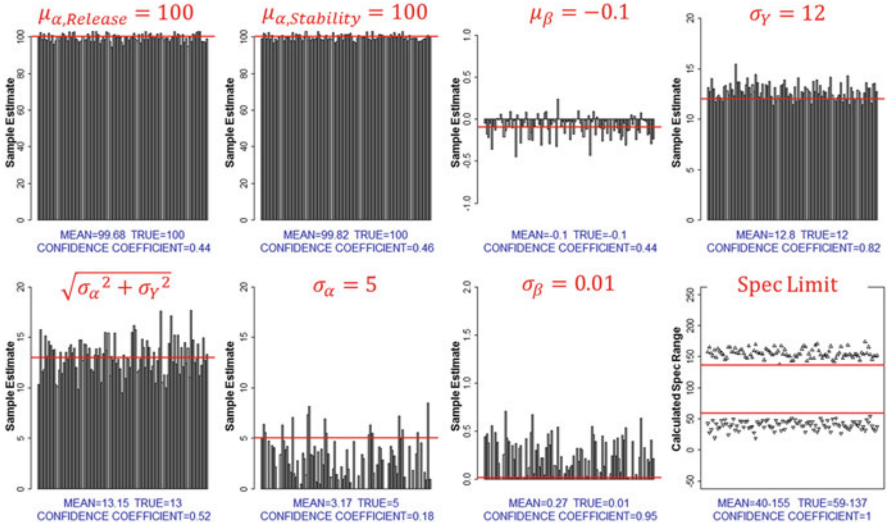


Fig. 8 Bar graphs of parameter estimates and resulting shelf-life specification limits from 10,000 simulations for $n_{stability} = 20$ and $\sigma_Y = 12$*Note:* x-axis is the index for 100 repeats of simulation; *red reference lines* are the TRUE parameter values

Table 1 Summary of mean parameter estimates and confidence coefficients using $\sigma_Y = 3$

	Parameter	μ_α	σ_α	μ_β	σ_β	σ_Y	Shelf-Life specification limits
# Stability lots	TRUE VALUE	100	5	−0.10	0.01	3	80–115
3	Mean estimate	100.18	2.63	−0.11	0.19	5.48	70–125
	Confidence coefficient	0.54	0.10	0.51	0.93	1.00	0.88
20	Mean estimate	99.95	1.39	−0.1	0.12	5.71	72–124
	Confidence coefficient	0.46	0.00	0.48	0.90	1.00	1.00

Table 2 Summary of mean parameter estimates and confidence coefficients using $\sigma_Y = 12$

	Parameter	μ_α	σ_α	μ_β	σ_β	σ_Y	Shelf-life specification limits
# Stability lots	TRUE VALUE	100	5	−0.10	0.01	12	59–137
3	Mean estimate	100.34	5.80	−0.12	0.44	12.17	36–158
	Confidence coefficient	0.54	0.45	0.49	0.93	0.57	0.88
20	Mean estimate	99.82	3.17	−0.10	0.27	12.80	40–155
	Confidence coefficient	0.46	0.18	0.44	0.95	0.82	1.00

$\sigma_Y = 12$, respectively.

$$[LSL,\ USL]_{true} = \mu_\alpha + \mu_\beta * t_{expiry} \mp 3 * \sqrt{\sigma_\alpha^2 + \left(t_{expiry} * \sigma_\beta\right)^2 + \sigma_Y^2} \tag{4}$$

The point estimates of the population parameters in the 100 repetitions of Figs. 5, 6, 7, and 8 are examined. The fixed mean intercept estimates, either from release data ($\hat{\mu}_R$) or stability data ($\hat{\mu}_\alpha$), generally approximate well the true population value. The variance components estimates are however biased, i.e., σ_α is underestimated while σ_β and σ_Y are overestimated. The cause of the systematic bias is still not clear to the author. One possible explanation is that the generated hypothetical data sets were generated using uncorrelated covariance structure for σ_α, σ_β, and σ_Y. The lmer R package however uses a correlated covariance structure to fit RCM. Further, there are some instances of having a zero result for the variance estimates due to negative variance being bounded to zero, as sometimes encountered when using mixed effects model. The bias and/or having zero-bounds in the variance parameter will diminish the reliability of the calculated specification limits using the simulation method.

The effects of the analytical variability and stability sample size are evaluated. For $\sigma_Y = 3$ (Figs. 5 and 6, Table 1), there is general improvement in precision of parameter estimates as lot sample size is increased from 3 to 20. The benefits of increased lot sample size are further amplified for $\sigma_Y = 12$ (Figs. 6 and 8, Table 2). Since the true fixed slope is $\mu_\beta = -0.1$, getting estimates that correctly estimates the negative direction and the magnitude of the parameter is crucial in setting specifications. Using only three lots (Fig. 7), there are several instances where incorrect positive fixed slopes are obtained. Of the instances that correctly estimated negative slopes, most severely overestimated the true -0.1 slope. Using 20 lots (Fig. 8), there are a lot more instances that negative direction is correctly estimated with slope magnitudes closer to the true -0.1 when compared to three lots. Despite having conservatively wider limits with only three lots, 12 of 100 repetitions failed to correctly contain the true shelf-life specification limits (i.e., confidence coefficient = 0.88, Fig. 7). In real practice, such inadequately set specification limits increases the likelihood getting out-of-specifications (OOS) results. Using 20 lots, the shelf-life specification limits more closely approximate the true limits and the confidence coefficient exceeds the nominal 0.95 target. In real practice, an adequately set specification limits minimizes the occurrences of OOS while serving as a practically useful manufacturing control strategy.

3.4 Benchmarking the Calculated Specification Limits Using Hierarchical Linear Modeling Simulation (HLMS) vs. an Industrial Practice (ADG)

The fixed effect ANCOVA analysis method described in Allen et al. (1991) (referred herein as ADG) calculates internal release limits given shelf-life specification limits and stability data. It is commonly practiced within the industry to adapt the ADG method by flipping the direction of calculation. Given release and stability data and shelf-life expiry (set *a priori* independently from stability analyses, e.g., through management business decision or alignment with other similar products), the specification limits are to be calculated. The adaptation of the method is formulated into Eq. (5). First, a base value is set from the release data either as a lower or upper 95 % confidence/99 % proportion tolerance limit ($TI_{Release}$), depending on whether product attribute is expected to decrease or increase, respectively. Second, the mean change over the shelf-life expiry $\left(\hat{b} * t_{expiry}\right)$ is accounted assuming stability lots have poolable common slope $\hat{b}$. Finally, the combined uncertainty in estimating the common slope and random variation $\left(t_{\gamma,df}\sqrt{\left(SE_{\hat{b}} * t_{expiry}\right)^2 + RMSE^2}\right)$ is accounted. If for example an attribute is expected to decrease, the $\hat{b}$ and $SE_{\hat{b}}$ terms in the *USL* formula are zeroed leaving only analytical variability added to the base value. For an attribute expected to increase, the *LSL* formula is analogously adjusted.

The calculated specification limits using the Hierarchical Linear Modeling Simulation (HLMS) are compared to those calculated using ADG. The results for the four $\left(n_{stability} = 3 \;\; and \;\; 20 \;\; lots\right)$ by $\left(\sigma_Y \; = \; 3 \;\; and \;\; 12\right)$ permutations are overlayed in Fig. 9 with the true specification limits calculated from Eq. (5) superimposed as red reference lines. The percent of the 100 repetitions wherein the proposed HLMS method yields narrower lower and upper specification limits than ADG as well as the confidence coefficients for each method are annotated in the plot and summarized in Table 3.

$$\begin{aligned} LSL_{ADG} &= lower\ TI_{Release} - |\hat{b}| * t_{expiry} - t_{\gamma,df}\sqrt{\left(SE_{\hat{b}} * t_{expiry}\right)^2 + RMSE^2} \\ USL_{ADG} &= upper\ TI_{Release} + |\hat{b}| * t_{expiry} + t_{\gamma,df}\sqrt{\left(SE_{\hat{b}} * t_{expiry}\right)^2 + RMSE^2} \end{aligned} \tag{5}$$

For all four $n_{stability}$ and σ_Y scenarios, majority of the 100 repetitions has HLMS limits being narrower than ADG. For example at $n_{stability} = 3$ and $\sigma_Y = 12$, HLMS has 83 % having larger calculated LSL and 75 % having smaller calculated USL than ADG (Table 3). Put differently, ADG is consistently wider than HLMS. This is expected as ADG is considered a conservative approach because it accounts for uncertainty (risk) as a worst-case scenario. It is worst-case because by using the release tolerance limit, the smallest (or largest) attribute measurement expected in the future is presupposed. Note that the release tolerance limit already incorporates

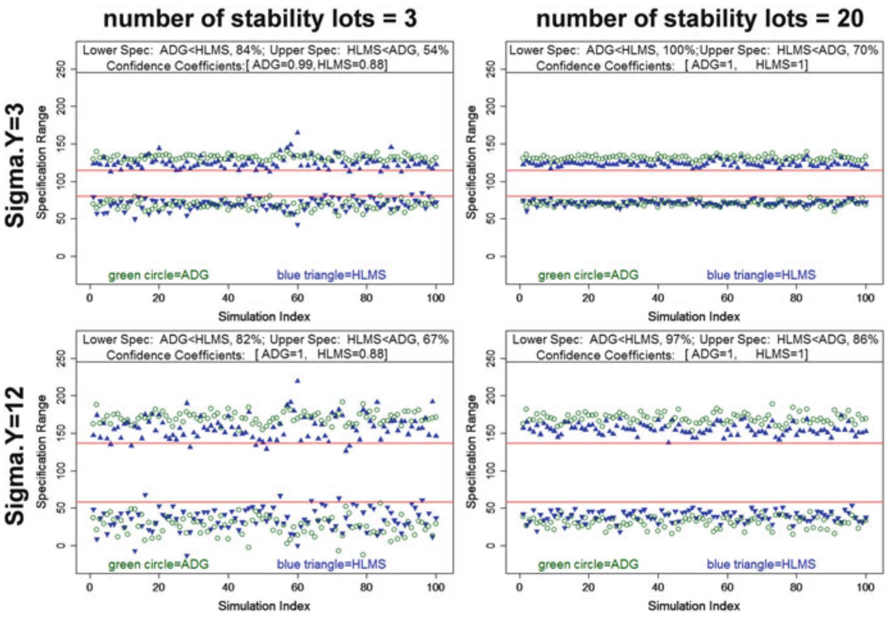


Fig. 9 Overlay plots comparing calculated specification limits using Allen et al. (1991) (ADG, *green circle*) and hierarchical linear model simulation (HLMS, *blue triangle*), relative to true limits (*red line*) for $n_{stability} = 3$ (left) vs. 20 (right) and for $\sigma_Y = 3$ (*top*) vs. 12 (*bottom*)

Table 3 Summary of relative comparison of width of specification limits and of the confidence coefficients for HLMS and ADG

		Percent of 100 repetitions where HLMS specification is narrower than ADG		Confidence coefficient	
# Stability lots	σ_Y	*LSL*	*USL*	ADG	HLMS
3	3	84	54	0.99	0.88
20	3	100	70	1.00	1.00
3	12	82	67	1.00	0.88
20	12	97	86	1.00	1.00

both process (lot-to-lot) and analytical (within-lot) variability. Additionally, RMSE is further incorporated (under the radical sign of Eq. 5) thus doubly accounting for analytical variability. The net effect is having conservatively wide specification limits using the ADG method.

Along with the relative widths of the calculated specification limits, the confidence coefficients are also evaluated in the comparison of HLMS vs. ADG. Due to the conservatively wide specification limits of ADG, its confidence coefficient more than adequately met the nominal target 0.95 with values almost being 1 in all scenarios. For the HLMS method at $n_{stability} = 3$, the RCM parameter estimates are not reliable or precise enough such that its confidence coefficient falls way below the nominal target of 0.95 (0.75 and 0.72 for $\sigma_Y = 3 \quad and \quad 12$, respectively, Table 3). With small stability sample sizes representative of the case for new biological entity at the time of regulatory filing, there is a large uncertainty in both the ANCOVA and RCM estimates. To minimize the probability of incurring future out-of-specification (OOS) values due to inadequately set specification limits, the conservative ADG approach may be the suitable method to apply. At $n_{stability} = 20$, the RCM parameter estimates are more reliable so the calculated specification limits approximate the true limits more closely while exceeding the nominal 0.95 confidence coefficient (0.97 and 1.00 for $\sigma_Y = 3 \quad and \quad 12$, respectively, Table 3). When stability sample sizes are larger at the time of regulatory filing as may be the case for biosimilar development, the results show that HLMS has superior performance than ADG.

4 Discussion

This paper highlights the opportunities for statistical applications in biosimilar product development. The first opportunity is related to the statistical framework for the analytical similarity assessment of structural and functional attributes between reference and biosimilar products. This aspect is foundational to the stepwise approach to provide a totality of evidence for demonstrating biosimilarity. The challenge is to formulate statistical methods that meaningfully assess biosimilarity while taking into account the practical issues of product sourcing and testing restrictions. It is a seminal area that will define the regulatory guidance for biosimilars.

The second opportunity related to typical Chemistry, Manufacturing and Control expectations for biologics license applications such as specification settings. The relatively larger number of biosimilar product lots characterized enables the use of more complex statistical models that account for random lot effects. Only a modest increase in code programming is entailed by switching from the fixed effects ANCOVA model used in Q1E as adapted from Allen et al. (1991) to the proposed hierarchical linear model simulation method. The intermediate calculation steps associated with the Q1E analyses (e.g., lot pooling, arbitrary lot subsetting if not all lots are poolable, deciding how to treat significant vs. non-significant slopes, determining what is the expected stability trend, etc.) make it cumbersome to implement. In contrast, all these intermediate steps are eliminated in the proposed simulation approach. There is no expectation that lots are poolable because the heterogeneity in intercepts and slopes are captured in the parameter estimates. Whether the fixed slope is significant or not, the magnitude of the slope standard deviation estimates will drive the prevailing temporal trend for the simulated future lots. Hence ultimately, the approach is a simpler concept and lends itself to easier automation for the establishment of specifications for multiple biosimilar product quality attributes.

There are some potential issues in using the hierarchical linear simulation method. One is that the simulation method is based on parametric model and the underlying distributions of parameters could impact the estimation of specification limits in practice. The validity of the results of mixed effects model is also influenced by the quality of the data being fitted. The results in this paper show that the variance components estimates using the R lmer function are systematically biased. The variance components estimates also sometimes come up negative which has to be bounded to zero. Either or both of these observations directly impact the simulated future values at shelf-life and therefore the reliability of the calculated shelf-life specification limits. The possible causes of this bias are being investigated further. The applicability of Bayesian approach to the simulation method is also being evaluated.

Overall, the hierarchical linear simulation method offers an improvement to the ADG method when stability sample size is large. The simulation has been used by Hospira, a Pfizer company for setting statistically-derived shelf-life specification limits submitted for actual biosimilar product candidates and have been received favorably by regulatory agencies.

References

FDA approves first biosimilar product Zarxio. March 6, 2015. Available at (http://www.fda.gov/newsevents/newsroom/pressannouncements/ucm436648.htm)

Scientific Considerations in Demonstrating Biosimilarity to a Reference Product: Guidance for Industry. April, 2015. Available at (http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM291128.pdf)

Tsong, Y., Shen, M., Dong, C. (2015), “Development of Statistical Approaches for Analytical Biosimilarity Evaluation”. Presentation at DIA Statistics Forum. April 2015. North Bethesda, MD

Burdick, R. K. and Ramírez, J. G. (2015) “Statistical Issues in Biosimilar Analytical Assessment: Perspectives on FDA ODAC Analysis, Presentation at DIA Conference, Washington, D. C., April.

Zhang, L. and Wu, S. (2015). “How to Set Biosimilarity Bounds in Biosimilar Developments”, Presentation at Joint Graybill and ICSA Annual Conference, Fort Collins, CO, June.

Chow, S.C. (2014). “On Assessment of Analytical Similarity in Biosimilar Studies”. Drug Des 3: 119. doi:10.4172/2169-0138.1000e124

Westlake, W. J. (1981). “Response to T. B. L. Kirkwood: Bioequivalence testing–a need to rethink. “Biometrics 37:589-594.

Schuirmann, D. J. (1987). “A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability.” J Pharmacokinet Biopharm 15(6): 657-680.

Q6A Specifications: Test Procedures and Acceptance Criteria for New Drug Substances and New Drug Products: Chemical Substances (2000). Available at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm134966.htm

Q6B Specifications: Test Procedures and Acceptance Criteria for Biotechnological / Biological Products (1999). Available at http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm073488.pdf

Q1A(R2) Stability Testing of New Drug Substances and Products (2003). Available at http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm073369.pdf

Q1E Evaluation of Stability Data (2004). Available at http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm073380.pdf

Allen, P. V., Dukes, G. R. and Gerger, M. E. (1991). "Determination of release limits: a general methodology." Pharm Res 8(9): 1210-1213.

Murphy, J. R. and Hofer J. D. (2002). "Establishing Shelf Life, Expiry Limits, and Release Limits." Drug Information Journal 36(4): 769-781.

Schofield, T., Apostol, I., Koeller, G., Powers, S., Stawicki, M. and Wolfe, R. A. (2008). "A Rational Approach for Setting and Maintaining Specifications for Biological and Biotechnology-Derived Products - Part 2." Biopharm International 21(7).

Dong, X., Tsong, Y. and Shen, M. (2014). "Statistical Considerations in Setting Product Specifications." Journal of Biopharmaceutical Statistics 25(2): 280-294.

Gelman, A. and Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. New York, NY, Cambridge University Press.

Part VII
Statistical Learning Methods and Applications with Large-Scale Data

A Statistical Method for Change-Set Analysis

Pei-Sheng Lin, Jun Zhu, Shu-Fu Kuo, and Katherine Curtis

Abstract In many scientific studies, it is of interest to group spatial units on a lattice with similar characteristics within a group but with distinction among groups. Here we develop a novel change-set method for this purpose, as a substantive extension of the existing change-point analysis for one-dimensional data in space or time. Our method addresses unique challenges resulting from the multi-dimensional space such as changes that occur abruptly in space and change sets of arbitrary shapes. In particular, we propose an entropy measure and establish quasi-likelihood estimation that accounts for covariates via change-set regression and spatial correlation via working covariance. For illustration, our method is applied to analyze a county-based socio-economic data set.

Keywords Change-set analysis • Entropy measure • Estimating equations • Quasi-likelihood estimation • Spatial lattice • Spatial statistics

1 Introduction

In many scientific studies, it is of interest and importance to group spatial units on a lattice to have similar characteristics within a group and distinction among groups. The motivating data example is from a research study aimed at discerning spatial patterns of poverty. A specific question of interest is to quantify similarities and

P.-S. Lin (✉) • S.-F. Kuo
Division of Biostatistics and Bioinformatics, National Health Research Institutes, Zhunan Township, Miaoli County, Taiwan

Department of Mathematics, National Chung Cheng University, Minxiong, Chiayi County, Taiwan
e-mail: pslin@nhri.org.tw

J. Zhu
Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA

K. Curtis
Department of Community and Environmental Sociology, University of Wisconsin-Madison, Madison, WI, USA

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_21

discrepancies in poverty rates among counties while accounting for socio-economic factors and spatial noise. In this paper, we develop a novel change-set method that can be applied to address such questions by grouping spatial units like counties into one or more change sets in the presence of covariates and spatial dependence.

In particular, we consider a change-set regression model such that the coefficient marking the change has the same absolute value but opposite signs for the true change set and its complement set. We then devise a method for change-set identification based on this model. A number of statistical issues need to be addressed. One, the true change set is unknown and there are many possibilities to consider as candidate change sets. Two, the parameters in the change-set regression models are not always identifiable. Three, the presence of spatial correlation complicates the estimation of parameters, which is otherwise relatively straightforward for the traditional change-point analysis (Lin 2011). Four, even with a correctly specified likelihood function, the computational burden can be substantial, making the estimation procedure infeasible (Song 2007).

To address these challenges, we develop a quasi-likelihood (QL) approach to parameter estimation and change-set identification. First, two sets of estimating equations are derived, one for the regression coefficients from the QL function and the other for the spatial covariance parameters by weighted least squares. Next, to identify the change sets, we propose a novel entropy measure in the form of a weighted sum of squared differences between the estimated mean functions from two change-set regression models, one for the response in the change set and the other in its complement set. A two-step procedure is devised to iteratively update the parameter estimates and change-set identification. Since the entropy measure has a quadratic form, the computation achieves convergence fairly quickly.

Compared with the edge detection methods for image restoration, our approach seems to be more general, as it does not require a regular grid of pixels and allows non-adjacent spatial units to belong to the same change set (Qiu 2005). Change-set analysis can also be related to cluster/classification analysis (Raftery 1994). However, the traditional classification methods, such as the *k*-means method (Hartigan 1975), or the spatial clustering methods, such as spatial scan statistics (Kulldorff 1997), generally do not take into account covariates or geographical locations. Our approach is more comprehensive in the sense that both covariates and spatial correlation are accounted for Jung (2009) and Zhang and Lin (2009). Finally, unlike Bayesian hierarchical modeling for spatial clustering, where relatively smooth changes are expected among nearby spatial units and intensive computation is often involved for inference (see, e.g., Lawson and Clark 2002; Gangnon and Clayton 2004), our method accommodates more abrupt changes in space and clusters of arbitrary shapes, while the computation is more feasible.

2 Model and Estimation

2.1 Change-Set Regression Model

Let $D \subset \mathbb{R}^2$ denote a spatial domain of interest. Suppose there are n observations of a response variable at sampling sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in D$. A sampling site could be the representative point in a spatial unit such as the location of a county seat or the centroid of an image pixel. Let $Y_i = Y(\boldsymbol{s}_i)$ denote the response variable, $\boldsymbol{x}_i = \boldsymbol{x}(\boldsymbol{s}_i) = (x_{i,1}, \ldots, x_{i,q})'$ denote q non-constant covariates, and $\theta_i = \theta(\boldsymbol{s}_i) = E(Y_i)$ denote the mean response at sampling site, $\boldsymbol{s}_i$ for $i = 1, \ldots, n$.

We model the response variable by a stochastic process $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$. Let $\mathcal{C}$ denote a change set such that the mean function of $Y(\boldsymbol{s})$ is $\mu_1(\boldsymbol{s})$ for $\boldsymbol{s} \in \mathcal{C}$, but is $\mu_0(\boldsymbol{s})$ for $\boldsymbol{s} \in \mathcal{C}^c$, where $\mathcal{C}^c = D - \mathcal{C}$ denotes the complement set of $\mathcal{C}$. The change-set regression model is

$$\theta(\boldsymbol{s}) = E\{Y(\boldsymbol{s})\} = \mu_0(\boldsymbol{s})I[\boldsymbol{s} \in \mathcal{C}^c] + \mu_1(\boldsymbol{s})I[\boldsymbol{s} \in \mathcal{C}], \tag{1}$$

where $I[\cdot]$ denotes the indicator function and the link function at $\boldsymbol{s}_i$ is

$$g\{\mu_0(\boldsymbol{s}_i)\} = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} \quad \text{and} \quad g\{\mu_1(\boldsymbol{s}_i)\} = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + \xi, \tag{2}$$

where β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$ are q regression coefficients (or, slopes) associated with the covariates, and ξ is a "jump coefficient" associated with the change set.

Let $\delta_i = I[\boldsymbol{s}_i \in \mathcal{C}]$ denote a "status variable" indicating whether site $\boldsymbol{s}_i$ belongs to the change set $\mathcal{C}$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ denote the response vector and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ denote the design matrix. Also let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ denote the mean response vector, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$ denote the true status vector, and $\boldsymbol{\delta}^c = 1 - \boldsymbol{\delta} = (\delta_1^c, \ldots, \delta_n^c)'$ denote the complement of the status vector $\boldsymbol{\delta}$. It can be easily verified that $\boldsymbol{\theta}$ can be re-expressed as either $\boldsymbol{\theta}_\delta = g^{-1}(\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \xi\boldsymbol{\delta})$ or $\boldsymbol{\theta}_{\delta^c} = g^{-1}\{(\beta_0 + \xi) + \boldsymbol{X}\boldsymbol{\beta} - \xi\boldsymbol{\delta}^c\}$. Further, we let $\boldsymbol{V}(\xi, \beta_0, \boldsymbol{\beta}; \boldsymbol{\tau}) = \operatorname{var}(\boldsymbol{Y})$ denote the covariance matrix of the response vector $\boldsymbol{Y}$ with covariance parameters $\boldsymbol{\tau}$.

To draw inference about the true but unknown change set, let T denote a candidate change set and with $w_i = I[\boldsymbol{s}_i \in T]$, let $\boldsymbol{w} = (w_1, \ldots, w_n)'$ denote the status vector corresponding to T. The mean response vector $\boldsymbol{\theta}$ written in terms of $\boldsymbol{w}$ is

$$\boldsymbol{\theta}_w \equiv \boldsymbol{\theta}_w(\xi_w, \beta_{0,w}, \boldsymbol{\beta}_w) = g^{-1}(\beta_{0,w} + \boldsymbol{X}\boldsymbol{\beta}_w + \xi_w\boldsymbol{w}), \tag{3}$$

where ξ_w, $\beta_{0,w}$, and $\boldsymbol{\beta}_w$ are the jump coefficient, intercept, and slopes. Analogously, for the complement set $T^c = D - T$, we let $\boldsymbol{w}^c = 1 - \boldsymbol{w}$ denote the complement status vector. Then, the mean response vector $\boldsymbol{\theta}$ can also be written in terms of $\boldsymbol{w}^c$ as

$$\boldsymbol{\theta}_{w^c} \equiv \boldsymbol{\theta}_{w^c}(\xi_{w^c}, \beta_{0,w^c}, \boldsymbol{\beta}_{w^c}) = g^{-1}(\beta_{0,w^c} + \boldsymbol{X}\boldsymbol{\beta}_{w^c} + \xi_{w^c}\boldsymbol{w}^c), \tag{4}$$

where ξ_{w^c}, $\beta_{0,w^c} = \beta_{0,w} + \xi_w$, and $\boldsymbol{\beta}_{w^c}$ are the complement jump coefficient, complement intercept, but the same slopes as in (3).

2.2 Estimating Equations

We now develop quasi-likelihood (QL) estimates of the jump coefficient ξ_w, the intercept $\beta_{0,w}$, and slopes $\boldsymbol{\beta}_w$ in model (3) for $\boldsymbol{\theta}_w$, given the covariance parameter $\boldsymbol{\tau}$ and a particular $\boldsymbol{w}$. Let $\boldsymbol{D}_{\xi_w,\beta_{0,w},\boldsymbol{\beta}_w}$ denote the first-order derivative matrix of $\boldsymbol{\theta}_w$ with respect to ξ_w, $\beta_{0,w}$, and $\boldsymbol{\beta}_w$. We define a QL score function for ξ_w, $\beta_{0,w}$, and $\boldsymbol{\beta}_w$ as

$$\boldsymbol{Q}(\xi_w, \beta_{0,w}, \boldsymbol{\beta}_w; \boldsymbol{\tau}) = \boldsymbol{D}'_{\xi_w,\beta_{0,w},\boldsymbol{\beta}_w}\boldsymbol{V}_w^{-1}(\boldsymbol{Y} - \boldsymbol{\theta}_w), \tag{5}$$

where $\boldsymbol{V}_w \equiv \boldsymbol{V}(\xi_w, \beta_{0,w}, \boldsymbol{\beta}_w; \boldsymbol{\tau})$ denotes the covariance matrix of $\boldsymbol{Y}$ given $\boldsymbol{\tau}$ and $\boldsymbol{w}$. Then we define the QL estimating equation as

$$\boldsymbol{Q}(\xi_w, \beta_{0,w}, \boldsymbol{\beta}_w; \boldsymbol{\tau})|_{(\xi_w,\beta_{0,w},\boldsymbol{\beta}_w)=(\hat{\xi}_w,\hat{\beta}_{0,w},\hat{\boldsymbol{\beta}}_w)} = \boldsymbol{0}. \tag{6}$$

The solutions $\hat{\xi}_w$, $\hat{\beta}_{0,w}$, and $\hat{\boldsymbol{\beta}}_w$ to (6) will be referred to as the QL estimates for the parameters in the mean function $\boldsymbol{\theta}_w$. When $\boldsymbol{Q}(\xi_w, \beta_{0,w}, \boldsymbol{\beta}_w; \boldsymbol{\tau})$ is nonlinear, a Newton-Raphson algorithm can be applied to solve (6). In particular, at the kth iteration for $k = 1, 2, \ldots$, we update $\left(\hat{\xi}_w^{(k)}, \hat{\beta}_{0,w}^{(k)}, \hat{\boldsymbol{\beta}}_w^{(k)\prime}\right)'$ by

$$\begin{aligned}\left(\hat{\xi}_w^{(k+1)}, \hat{\beta}_{0,w}^{(k+1)}, \hat{\boldsymbol{\beta}}_w^{(k+1)\prime}\right)' &= \left(\hat{\xi}_w^{(k)}, \hat{\beta}_{0,w}^{(k)}, \hat{\boldsymbol{\beta}}_w^{(k)\prime}\right)' \\ &+ \left(\boldsymbol{D}'_{\xi_w,\beta_{0,w},\boldsymbol{\beta}_w}\boldsymbol{V}_w^{-1}\boldsymbol{D}_{\xi_w,\beta_{0,w},\boldsymbol{\beta}_w}\right)^{-1}\boldsymbol{D}'_{\xi_w,\beta_{0,w},\boldsymbol{\beta}_w}\boldsymbol{V}_w^{-1}(\boldsymbol{Y} - \boldsymbol{\theta}_w)\Big|_{(\xi_w,\beta_{0,w},\boldsymbol{\beta}'_w)'=\left(\hat{\xi}_w^{(k)},\hat{\beta}_{0,w}^{(k)},\hat{\boldsymbol{\beta}}_w^{(k)\prime}\right)'}.\end{aligned}$$

Next, we develop QL estimates of the complement jump coefficient ξ_{w^c}, the complement intercept β_{0,w^c}, and complement slopes $\boldsymbol{\beta}_{w^c}$ in model (4) for $\boldsymbol{\theta}_{w^c}$. Since $\beta_{0,w^c} = \beta_{0,w} + \xi_w$, we estimate the complement intercept β_{0,w^c} by $\hat{\beta}_{0,w^c} = \hat{\beta}_{0,w} + \hat{\xi}_w$ to ensure the same baseline. Let $\boldsymbol{\theta}^*_{w^c} = g^{-1}\{\hat{\beta}_{0,w^c} + \boldsymbol{X}\boldsymbol{\beta}_{w^c} + \xi_{w^c}\boldsymbol{w}^c\}$, $\boldsymbol{V}^*_{w^c} = \boldsymbol{V}(\xi_{w^c}, \hat{\beta}_{0,w^c}, \boldsymbol{\beta}_{w^c}; \boldsymbol{\tau})$, and $\boldsymbol{Q}^*(\xi_{w^c}, \boldsymbol{\beta}_{w^c}; \boldsymbol{\tau}) = \boldsymbol{D}'_{\xi_{w^c},\boldsymbol{\beta}_{w^c}}\boldsymbol{V}^{*-1}_{w^c}(\boldsymbol{Y} - \boldsymbol{\theta}^*_{w^c})$. Given $\boldsymbol{\tau}$ and $\boldsymbol{w}^c$, we estimate ξ_{w^c} and $\boldsymbol{\beta}_{w^c}$ by solving the QL estimating equation

$$\boldsymbol{Q}^*(\xi_{w^c}, \boldsymbol{\beta}_{w^c}; \boldsymbol{\tau})|_{(\xi_{w^c}, \boldsymbol{\beta}_{w^c})=(\hat{\xi}_{w^c}, \hat{\boldsymbol{\beta}}_{w^c})} = \mathbf{0}. \tag{7}$$

The solutions $\hat{\xi}_{w^c}$ and $\hat{\boldsymbol{\beta}}_{w^c}$ to (7) can be computed analogously to (6) and will be referred to as the QL estimates for ξ_{w^c} and $\boldsymbol{\beta}_{w^c}$ in the mean function $\boldsymbol{\theta}_{w^c}$.

Last, to estimate the covariance parameters $\boldsymbol{\tau}$, we construct a third estimating equation. For a given status vector $\boldsymbol{w}$, let $\boldsymbol{V}^w(\boldsymbol{\tau}) \equiv \boldsymbol{V}_w(\hat{\xi}_w, \hat{\beta}_{0,w}, \hat{\boldsymbol{\beta}}_w; \boldsymbol{\tau})$ denote the covariance matrix of $\boldsymbol{Y}$ given $\hat{\xi}_w$, $\hat{\beta}_{0,w}$, and $\hat{\boldsymbol{\beta}}_w$. Let vec($\boldsymbol{A}$) denote the vectorization of matrix $\boldsymbol{A}$. We define $\boldsymbol{Z}_w = \text{vec}\left\{(\boldsymbol{Y} - \hat{\boldsymbol{\theta}}_w)(\boldsymbol{Y} - \hat{\boldsymbol{\theta}}_w)'\right\}$ and $\boldsymbol{\Omega}_w(\boldsymbol{\tau}) = \text{vec}\{\boldsymbol{V}^w(\boldsymbol{\tau})\}$, where $\hat{\boldsymbol{\theta}}_w = g^{-1}(\hat{\beta}_{0,w} + \boldsymbol{X}\hat{\boldsymbol{\beta}}_w + \hat{\xi}_w \boldsymbol{w})$. We then obtain an estimate $\hat{\boldsymbol{\tau}}$ for $\boldsymbol{\tau}$ by least squares:

$$\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} \{\boldsymbol{Z}_w - \boldsymbol{\Omega}_w(\boldsymbol{\tau})\}'\{\boldsymbol{Z}_w - \boldsymbol{\Omega}_w(\boldsymbol{\tau})\}. \tag{8}$$

3 Change-Set Identification Method

3.1 Single Change-Set Identification

We start with identification of a single change set and for a given status vector $\boldsymbol{w}$, we define a QL entropy measure as

$$S(\xi_w, \xi_{w^c}; \boldsymbol{w}) = n^{-1}(\boldsymbol{\theta}_w - \boldsymbol{\theta}_{w^c})' \left(\boldsymbol{V}_w^{-1} + \boldsymbol{V}_{w^c}^{-1}\right) (\boldsymbol{\theta}_w - \boldsymbol{\theta}_{w^c}), \tag{9}$$

which involves only the mean and covariance functions of the responses. Let $\mathcal{W}$ denote the collection of all possible $\boldsymbol{w}$'s. Let $\hat{\boldsymbol{\theta}}_w = g^{-1}(\hat{\beta}_{0,w} + \boldsymbol{X}\hat{\boldsymbol{\beta}}_w + \hat{\xi}_w \boldsymbol{w})$ and $\hat{\boldsymbol{\theta}}_{w^c} = g^{-1}(\hat{\beta}_{0,w^c} + \boldsymbol{X}\hat{\boldsymbol{\beta}}_{w^c} + \hat{\xi}_{w^c} \boldsymbol{w}^c)$ denote the estimated mean functions, where $\hat{\beta}_{0,w^c} = \hat{\beta}_{0,w} + \hat{\xi}_w$. Let $\hat{\boldsymbol{V}}_w = \boldsymbol{V}(\hat{\xi}_w, \hat{\beta}_{0,w}, \hat{\boldsymbol{\beta}}_w; \hat{\boldsymbol{\tau}})$ and $\hat{\boldsymbol{V}}_{w^c} = \boldsymbol{V}(\hat{\xi}_{w^c}, \hat{\beta}_{0,w^c}, \hat{\boldsymbol{\beta}}_{w^c}; \hat{\boldsymbol{\tau}})$ denote the estimated covariance matrices. We estimate the true status vector $\boldsymbol{\delta}$ by maximizing the difference between the estimated mean functions $\hat{\boldsymbol{\theta}}_w$ and $\hat{\boldsymbol{\theta}}_{w^c}$ weighted by the estimated covariance functions, for all $\boldsymbol{w} \in \mathcal{W}$. That is, with $S(\hat{\xi}_w, \hat{\xi}_{w^c}; \boldsymbol{w}) = n^{-1}(\hat{\boldsymbol{\theta}}_w - \hat{\boldsymbol{\theta}}_{w^c})' \left(\hat{\boldsymbol{V}}_w^{-1} + \hat{\boldsymbol{V}}_{w^c}^{-1}\right) (\hat{\boldsymbol{\theta}}_w - \hat{\boldsymbol{\theta}}_{w^c})$, we estimate the status vector by

$$\hat{\boldsymbol{\delta}} = \arg\max_{\boldsymbol{w} \in \mathcal{W}} S(\hat{\xi}_w, \hat{\xi}_{w^c}; \boldsymbol{w}). \tag{10}$$

We then develop a two-step procedure that toggles between the estimation step (6)–(8) and the identification step (10) for identifying a single change set. For parameter identification and numerical stability, we fix $\beta_{0,w}$ and $\boldsymbol{\tau}$ for all $\boldsymbol{w} \in \mathcal{W}$. Further, to distinguish the two steps, we use $\tilde{\cdot}$ to denote the QL estimates for a specific status vector $\boldsymbol{w}$ in the estimation step and $\hat{\cdot}$ to denote the QL estimates for a general status vector $\boldsymbol{w}$ in the identification step. The computational algorithm is as follows.

Computational Algorithm

0. Initialization:
   Assuming independence (i.e., no spatial dependence), for each $\boldsymbol{w} \in \mathcal{W}$, obtain the initial QL estimates $\left(\hat{\xi}_w^{(0)}, \hat{\beta}_{0,w}^{(0)}, \hat{\boldsymbol{\beta}}_w^{(0)\prime}\right)^\prime$ from (6) and $\left(\hat{\xi}_{w^c}^{(0)}, \hat{\boldsymbol{\beta}}_{w^c}^{(0)\prime}\right)^\prime$ from (7). Evaluated at these estimates, (10) gives an initial status vector estimate $\hat{\boldsymbol{\delta}}_0$. Let $\hat{\boldsymbol{\tau}}^{(0)} \equiv \mathbf{0}$.
1. Parameter estimation update:
   Given $\hat{\boldsymbol{\delta}}_m$ and $\hat{\boldsymbol{\tau}}^{(m)}$ at iteration $m = 0, 1, 2, \ldots$, first estimate the jump coefficient $\xi_{\hat{\delta}_m}$, intercept $\beta_{0,\hat{\delta}_m}$, and slopes $\boldsymbol{\beta}_{\hat{\delta}_m}$ by (6) with $\boldsymbol{Q}\left(\xi_{\hat{\delta}_m}, \beta_{0,\hat{\delta}_m}, \boldsymbol{\beta}_{\hat{\delta}_m}; \hat{\boldsymbol{\tau}}^{(m)}\right)$ associated with $\hat{\boldsymbol{\delta}}_m$. Then estimate the complement jump coefficient $\xi_{\hat{\delta}_m^c}$ and slopes $\boldsymbol{\beta}_{\hat{\delta}_m^c}$ by (7) with $\boldsymbol{Q}^*\left(\xi_{\hat{\delta}_m^c}, \boldsymbol{\beta}_{\hat{\delta}_m^c}; \hat{\boldsymbol{\tau}}^{(m)}\right)$ associated with $\hat{\boldsymbol{\delta}}_m^c = 1 - \hat{\boldsymbol{\delta}}_m$. The updated QL estimates are denoted $\tilde{\xi}_{\hat{\delta}_m}$, $\tilde{\beta}_{0,\hat{\delta}_m}$, $\tilde{\boldsymbol{\beta}}_{\hat{\delta}_m}$, $\tilde{\xi}_{\hat{\delta}_m^c}$, and $\tilde{\boldsymbol{\beta}}_{\hat{\delta}_m^c}$. Next, update $\hat{\boldsymbol{\tau}}^{(m)}$ to $\hat{\boldsymbol{\tau}}^{(m+1)}$ by (8), where $\boldsymbol{Z}_{\hat{\delta}_m} = \text{vec}\left\{(\boldsymbol{Y} - \tilde{\boldsymbol{\theta}}_{\hat{\delta}_m})(\boldsymbol{Y} - \tilde{\boldsymbol{\theta}}_{\hat{\delta}_m})^\prime\right\}$, $\boldsymbol{\Sigma}_{\hat{\delta}_m}(\boldsymbol{\tau}) = \text{vec}\left\{\boldsymbol{V}_{\hat{\delta}_m}(\tilde{\xi}_{\hat{\delta}_m}, \tilde{\beta}_{0,\hat{\delta}_m}, \tilde{\boldsymbol{\beta}}_{\hat{\delta}_m}; \boldsymbol{\tau})\right\}$ and $\tilde{\boldsymbol{\theta}}_{\hat{\delta}_m} = g^{-1}(\tilde{\beta}_{0,\hat{\delta}_m} + \boldsymbol{X}\tilde{\boldsymbol{\beta}}_{\hat{\delta}_m} + \tilde{\xi}_{\hat{\delta}_m}\hat{\boldsymbol{\delta}}_m)$.
2. Status vector update:
   Given $\tilde{\beta}_{0,\hat{\delta}_m}$ and $\hat{\boldsymbol{\tau}}^{(m+1)}$, for each $\boldsymbol{w} \in \mathcal{W}$, obtain the parameter estimates $\hat{\xi}_w^{(m+1)}$ and $\hat{\boldsymbol{\beta}}_w^{(m+1)}$ by (6) with $\boldsymbol{Q}\left(\xi_w, \tilde{\beta}_{0,\hat{\delta}_m}, \boldsymbol{\beta}_w; \hat{\boldsymbol{\tau}}^{(m+1)}\right)$, and the parameter estimates $\hat{\xi}_{w^c}^{(m+1)}$ and $\hat{\boldsymbol{\beta}}_{w^c}^{(m+1)}$ by (7) with $\boldsymbol{Q}^*\left(\xi_{w^c}, \boldsymbol{\beta}_{w^c}; \hat{\boldsymbol{\tau}}^{(m+1)}\right)$. Then by (10), update the status vector to $\hat{\boldsymbol{\delta}}_{m+1}$. If $\hat{\boldsymbol{\delta}}_{m+1} \neq \hat{\boldsymbol{\delta}}_m$, then return to Step 1. Otherwise, convergence is achieved.

At convergence after say M iterations, we let $\hat{\boldsymbol{\delta}}_M \equiv \hat{\boldsymbol{\delta}}$ denote the final status vector estimate. Let $(\hat{\xi}_w^{(M)}, \hat{\boldsymbol{\beta}}_w^{(M)})$ denote the final QL estimates of $(\xi_w, \boldsymbol{\beta}_w)$ under $\boldsymbol{w}$, $(\hat{\xi}_{w^c}^{(M)}, \hat{\boldsymbol{\beta}}_{w^c}^{(M)})$ denote the final QL estimates of $(\xi_{w^c}, \boldsymbol{\beta}_{w^c})$ under $\boldsymbol{w}^c$, and $\tilde{\beta}_{0,\hat{\delta}_M}$ denote the final QL estimate of $\beta_{0,w}$ for all $\boldsymbol{w} \in \mathcal{W}$. Since $\hat{\boldsymbol{\beta}}_w$ and $\hat{\boldsymbol{\beta}}_{w^c}$ can be shown to be consistent for each $\boldsymbol{w} \in \mathcal{W}$ and within each iteration $\beta_{0,w}$ and $\boldsymbol{\tau}$ are fixed for all $\boldsymbol{w} \in \mathcal{W}$, the maximization in (10) becomes an estimation problem of ξ_w. Since (10) is quadratic, the convergence generally takes one or two iterations, as indicated in a simulation study (not shown here). Thus, the computation of the proposed QL entropy method is on the order of the size of $\mathcal{W}$ and is computational feasible, provided that $|\mathcal{W}|$ is not too big. We will refer to the above as an QL entropy method for a single change-set identification.

3.2 Multiple Change-Set Identification

Now, we consider the identification of multiple change sets sequentially and propose a search algorithm based on hypothesis testing. Recall that $\hat{\xi}_w^{(M)}$ and $\hat{\xi}_{w^c}^{(M)}$ are the final QL estimates of ξ_w under $\boldsymbol{w}$ and ξ_{w^c} under $\boldsymbol{w}^c$, respectively, for any given $\boldsymbol{w} \in \mathcal{W}$. We consider the test statistic $(\hat{\xi}_w - \hat{\xi}_{w^c})/\sigma_{\xi_w}$ and compare it with the standard normal distribution.

One way to estimate $\sigma^2_{\xi_w}$ is by Taylor's expansions of $\boldsymbol{Q}(\xi_w, \beta_{0,w}, \boldsymbol{\beta})$ and $\boldsymbol{Q}^*(\xi_{w^c}, \boldsymbol{\beta})$ with respect to ξ_w and ξ_{w^c}. Define $\boldsymbol{H} = \left(\boldsymbol{D}'_{\xi_w} \boldsymbol{V}_w^{-1} \boldsymbol{D}_{\xi_w}\right)^{-1} \boldsymbol{D}'_{\xi_w} + \left(\boldsymbol{D}'_{\xi_{w^c}} \boldsymbol{V}_{w^c}^{-1} \boldsymbol{D}_{\xi_{w^c}}\right)^{-1} \boldsymbol{D}'_{\xi_{w^c}}$, where $\boldsymbol{D}_{\xi_w}$ denotes the derivative matrix of $\boldsymbol{\theta}_w$ with respect to ξ_w, and $\boldsymbol{D}_{\xi_{w^c}}$ denotes the derivative matrix of $\boldsymbol{\theta}_{w^c}$ with respect to ξ_{w^c}. After some tedious calculation (not shown), we have $\sigma^2_{\xi_w} = \boldsymbol{H}\boldsymbol{V}_w^{-1}\boldsymbol{H}' + o(1)$ as $n \to \infty$. Thus, our estimate for $\sigma^2_{\xi_w}$ is

$$\hat{\sigma}^2_{\xi_w} = \hat{\boldsymbol{H}}\hat{\boldsymbol{V}}_w^{-1}\hat{\boldsymbol{H}}', \tag{11}$$

where $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{V}}_w$ are $\boldsymbol{H}$ and $\boldsymbol{V}_w$ evaluated at the final QL estimates.

We develop a sequential search algorithm as follows. For each candidate status vector $\boldsymbol{w} \in \mathcal{W}$, we compute the test statistic, $Z_{\xi_w} = \{\hat{\xi}_w^{(M)} - \hat{\xi}_{w^c}^{(M)}\}/\hat{\sigma}_{\xi_w}$, and compare Z_{ξ_w} with the standard normal distribution to obtain an approximate p-value. To adjust for multiple comparisons over $\mathcal{W}$, a collection of status vectors $\boldsymbol{w}$ with say N status vectors, we apply the false discovery rate (Banjamini and Hochberg 1995). Specifically, we put the p-values $p_{(1)}, \ldots, p_{(N)}$ in the ascending order and find the largest k^* such that $p_{(k^*)} \leq k^*\alpha/N$ for a given level of significance α. If there is no such k^*, then the search algorithm stops. Otherwise, let $\mathcal{W}^*$ denote an updated collection of status vectors $\boldsymbol{w}$ whose p-values are less than the threshold $k^*\alpha/N$. We then repeat the QL entropy method for a single change set in Sect. 3.1 but this time over $\mathcal{W}^*$ in search of the next change set.

4 Data Example

In a study to assess the social and economic factors of poverty in the US, census data in 535 counties of the five states in the upper Midwest (Illinois, Indiana, Michigan, Minnesota, and Wisconsin) were compiled (Curtis et al. 2013). For illustration of the methodology developed in Sects. 2 and 3, we focus on census data from the year 1960 and examined poverty in relation to industrial structures and socio-economic compositions. The response variable is a poverty rate, computed as the proportion of the county's population living below the poverty threshold. Let $\boldsymbol{s}_i$ denote the latitude and longitude and $Y_i^* \equiv Y^*(\boldsymbol{s}_i)$ denote the poverty rate in county i, for $i = 1, \ldots, n$, where $n = 535$. Figure 1a maps the county-level poverty rates with a range from

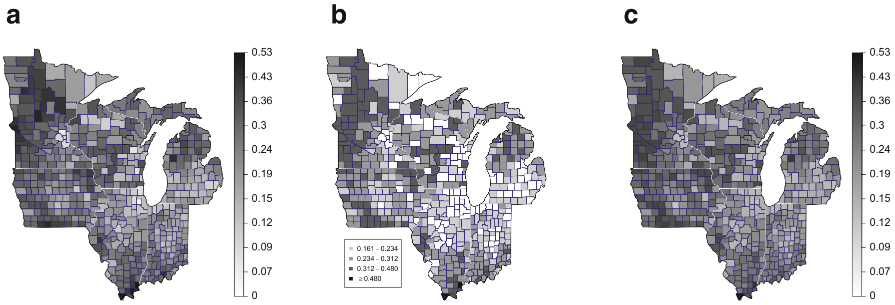


Fig. 1 Maps of (**a**) observed poverty rate, (**b**) estimated change sets, and (**c**) estimated poverty rate based on final change set model

0.055 to 0.526 and a mean value of 0.245. The simulation results suggested that Gaussian models are more reliable than binary models and thus, we transformed the poverty rate to the logistic scale $Y_i = \log\{Y_i^*/(1-Y_i^*)\}$ and let $\boldsymbol{Y} = (Y_1, \ldots, Y_{535})'$.

Change sets are considered to be due to unknown factors after incorporation of all the known factors and the signal-to-noise ratio. The number of possible status vectors is very large (2^{535}) and it is computationally infeasible to search the entire collection $\mathcal{W}$. Thus, we used the horizontal lines $Y_i^* = 0.05 + 0.001k$ to identify change sets similar to the simulation study, although the separating line was based on locations in the simulation and based on the response here. The change sets, if identified, are defined by the range of income and reflect unknown factors. Specifically, define the status variable for county i as $w_{k,i} = I[Y_i^* \geq 0.05 + 0.001k]$ which indicates whether county i had a poverty rate greater than $0.05 + 0.001k$. Then, $\boldsymbol{w}_k = (w_{k,1}, \ldots, w_{k,535})'$ is a status vector for the change set $Y_i^* \geq 0.05 + 0.001k$, where $k = 1, \ldots, 472$. We let $\mathcal{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{472}\}$ denote the collection of status vectors for the possible change sets.

The covariates representing the industrial structures are the proportions of the county population employed in five dominant industries, namely, agriculture (pag), mining (pex), manufacturing (pman), services (pserve), and FIRE (finance, insurance, and real estate) (pfire). In addition, the covariates representing an area's racial and ethnic compositions are proportions of four racial/ethnic groups, namely, white (pwh), African American (pblk), American Indian (pind), and Hispanic (phisp).

For variable selection among the covariates, we use a forward selection method with a quasi-deviance (QDEV) criterion (Lin 2011). Let $\boldsymbol{\beta}_{\tau_1}$ with q_1 elements and $\boldsymbol{\beta}_{\tau_2}$ with q_2 elements denote the parameter sets of two nested models under testing with $q_1 > q_2$. For correlated data, the QDEV function $D(\boldsymbol{\beta}_{\tau_1}, \boldsymbol{\beta}_{\tau_2})$, defined as

$$(1/2)\{\boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_1}) - \boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_2})\}' \left[\boldsymbol{V}^{-1}(\boldsymbol{\beta}_{\tau_1})\{\boldsymbol{Y} - \boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_1})\} + \boldsymbol{V}^{-1}(\boldsymbol{\beta}_{\tau_2})\{\boldsymbol{Y} - \boldsymbol{\theta}(\boldsymbol{\beta}_{\tau_2})\}\right]$$

can be used for model selection. Let $\hat{\boldsymbol{\beta}}_{\tau_1}$ and $\hat{\boldsymbol{\beta}}_{\tau_2}$ denote the QL estimates of $\boldsymbol{\beta}_{\tau_1}$ and $\boldsymbol{\beta}_{\tau_2}$, respectively. For nested models $\boldsymbol{\beta}_{\tau_1} \supset \boldsymbol{\beta}_{\tau_2}$, it holds that $2D(\hat{\boldsymbol{\beta}}_{\tau_1}, \hat{\boldsymbol{\beta}}_{\tau_2})$ converges in distribution to a chi-squared distribution $\chi^2_{q_0}$ with $q_0 = q_1 - q_2$ degrees of freedom. Variable selection can then be conducted by comparing the QDEV value with the chi-squared distribution. For example, the model with $\boldsymbol{\beta}_{\tau_1}$ is selected if $2D(\hat{\boldsymbol{\beta}}_{\tau_1}, \hat{\boldsymbol{\beta}}_{\tau_2}) > \chi^2_{q0,1-\alpha}$, where α is the level of significance and $\chi^2_{q0,1-\alpha}$ is the $(1-\alpha)$th quantile of the χ^2_{q0} distribution.

To find an integrated change-set model that accounts for covariates and spatial correlation, we first applied a QL method for spatial linear regression models without involving any change set. Five covariates, `pag`, `pman`, `pserve`, `pfire`, `pblk`, were selected using the QDEV-based forward selection. We then sequentially identified change sets based on the model $\theta_i^{\mathcal{C}} = \beta_0 + \beta_1 \texttt{pag}_i + \beta_2 \texttt{pman}_i + \beta_3 \texttt{pserve}_i + \beta_4 \texttt{pfire}_i + \beta_5 \texttt{pblk}_i + \xi_{\mathcal{C}} \delta^{\mathcal{C}}(\boldsymbol{s}_i)$, where β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_5)'$ are the regression coefficients, and $\xi_{\mathcal{C}}$ is the jump coefficient related to the change set $\mathcal{C}$. We applied the sequential search procedure shown in Sect. 3.2 to identify multiple change sets. With the procedure shown in Sect. 3, four change sets were identified, namely, $\mathcal{C}_1 = \{\boldsymbol{s}_i : Y_i \geq 0.480\}$, $\mathcal{C}_2 = \{\boldsymbol{s}_i : 0.312 \leq Y_i < 0.480\}$, $\mathcal{C}_3 = \{\boldsymbol{s}_i : 0.234 \leq Y_i < 0.312\}$, and $\mathcal{C}_4 = \{\boldsymbol{s}_i : 0.161 \leq Y_i < 0.234\}$ (Fig. 1b). Thus, the final model is

$$E(Y_i) = \beta_0 + \beta_1 \texttt{pag}_i + \beta_2 \texttt{pman}_i + \beta_3 \texttt{pserve}_i + \beta_4 \texttt{pfire}_i + \beta_5 \texttt{pblk}_i + \sum_{j=1}^{4} \xi_{\mathcal{C}_j} \delta^{\mathcal{C}_j}(\boldsymbol{s}_i). \tag{12}$$

We fitted the data by the model (12) and obtained the parameter estimates by applying the QL method shown in Sect. 2.2 (Table 1). In Table 1, we also computed the standard error for each estimated parameters by (11). To further assess the models with and without change sets, we computed a weighted least squares (WLS) error $(\boldsymbol{Y} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{V}}_{\boldsymbol{\tau}}^{-1}(\boldsymbol{Y} - \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{V}}_{\boldsymbol{\tau}}$ is an estimate of the covariance matrix for $\boldsymbol{Y}$.

Table 1 Parameter estimates with standard errors (in parenthesis) and p-values for the poverty data example based on all-covariate regression model (without change sets) and the final change-set model with covariates `pag`, `pman`, `pserve`, `pfire`, and `pblk` and four clusters $\mathcal{C}_1, \ldots, \mathcal{C}_4$

	No change sets		With change sets	
Covariate	Estimate	p-value	Estimate	p-value
Intercept	−1.6(0.13)	< 0.01	−1.7(0.08)	< 0.01
`pag`	2.2(0.13)	< 0.01	0.7(0.10)	< 0.01
`pman`	−0.6(0.13)	< 0.01	−0.4(0.09)	< 0.01
`pserve`	1.4(0.32)	< 0.01	0.5(0.23)	0.015
`pfire`	−6.5(0.93)	< 0.01	−4.8(0.65)	< 0.01
`pblk`	2.4(0.25)	< 0.01	1.4(0.19)	< 0.01
$\mathcal{C}_1$			1.1(0.08)	< 0.01
$\mathcal{C}_2$			0.9(0.03)	< 0.01
$\mathcal{C}_3$			0.6(0.02)	< 0.01
$\mathcal{C}_4$			0.3(0.02)	< 0.01
WLS error	763		160	

The weighted least squares (WLS) errors are also reported

Accounting for change sets in the model substantially improved the model fit based on the WLS error values. The regression coefficients and the jump coefficients for the four change sets are all significant. Four change sets are identified and the counties in the same change set tend to be near one another. The counties in the change-sets $\mathcal{C}_1$ and $\mathcal{C}_2$ have the higher poverty rates and tend to concentrate in the northernmost and southernmost counties of the region. Many counties along the shores of the Great Lakes are in the complement of the change sets with low poverty rates, in part due to the positive association with strong, stable manufacturing jobs in the area during the period of pre-deindustrialization. Further, for the region as a whole the proportions of agriculture and service employment were positively related to poverty, whereas those of manufacturing and FIRE had a negative relation. Generally, counties with strong ties to industries that are vulnerable to contraction are also vulnerable to higher rates of poverty. Findings suggest that counties more heavily dependent on agriculture and service had higher rates of poverty, whereas counties with strong ties to manufacturing and the FIRE sector had lower poverty rates. Finally, the estimated correlation is $\hat{\rho}_{i,j} = 0.75 \exp\left(-0.602\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2\right)$, with quite a strong spatial correlation, ranging from 0.61 to 0.75, for counties within 50 km of each other.

5 Conclusions and Discussion

Here we have developed a novel QL entropy method for change-set analysis, providing a substantive extension of the existing change-point analysis for one-dimensional data in space or time. An entropy measure has been proposed for identifying change sets and a quasi-likelihood procedure for change-set regression models has been devised to account for covariates and spatial correlation with feasible computation. For illustration, we have analyzed a county-level socio-economic data set and interesting spatial patterns of changes have been identified. A natural extension of our method would be the simultaneous identification of change sets in space and change points in time. This may be accomplished by creating spatio-temporal status vectors, which we leave for future investigation.

References

Banjamini, Y. and 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.

Curtis, K.J., Reyes, P.E., O'Connell, H., and Zhu, J. (2013). Assessing the spatial concentration and temporal persistence of poverty: industrial structure, racial/ethnic composition, and the complex links to poverty. *Spatial Demography* **1**, 178–194.

Gangnon, R.E. and Clayton, M.K. (2004). Likelihood-based tests for localized spatial clustering of disease. *Environmetrics* **15**, 797–810.

Hartigan, J.A. (1975). *Clustering Algorithm.* Wiley, New York.

Jung, I. (2009). A generalized linear models approach to spatial scan statistics for covariate adjustment. *Statistics in Medicine* **28**, 1131–1143.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods* **26**, 1487–1496.

Lawson, A.B. and Clark, A. (2002). Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine* **21**, 359–370.

Lin, P.-S. (2011). Quasi-deviance functions for spatially correlated data. *Statistica Sinica* **21**, 1785–1806.

Qiu, P. (2005). *Image Processing and Jump Regression Analysis*. Wiley, New Jersey.

Raftery, A.E. (1994). Change point and change curve modeling in stochastic processes and spatial statistics. *Journal of Applied Statistical Science* **1**, 403–424.

Song, P.X.K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer, New York.

Zhang, T. and Lin, G. (2009). Cluster detection based on spatial associations and iterated residuals in generalized linear mixed models. *Biometrics* **65**, 353–360.

An Alarm System for Flu Outbreaks Using Google Flu Trend Data

Gregory Vaughan, Robert Aseltine, Sy Han Chiou, and Jun Yan

Abstract Outbreaks of influenza pose a serious threat to communities and hospital resources. It is important for health care providers not only to know the seasonal trend of influenza, but also to be alarmed when unusual outbreaks occur as soon as possible for more efficient, proactive resource allocation. Google Flu Trends data showed a good match in trend patterns, albeit not in exact occurrences, with the proportion of physician visits attributed to influenza from the Centers for Disease Control, and, hence, provide a timely, inexpensive data source to develop an alarm system for outbreaks of influenza. For the State of Connecticut, using weekly Google Flu Trends data from 2003 to 2012, an exponentially weighted moving average control chart was developed after removing the seasonal trend from the observed data. The control chart was tested with the 2013–2015 data from the Center for Disease Control, and was able to issue an alarm at the unusually earlier outbreak in the 2012–2013 season.

Keywords Control chart • Exponentially weighted moving average process • Influenza • Statistical process control

G. Vaughan (✉)
Department of Statistics, University of Connecticut, Storrs, CT, USA
e-mail: gregory.vaughan@uconn.edu

R. Aseltine
Division of Behavioral Science and Community Health, University of Connecticut Health Center, Farmington, CT, USA

Center for Public Health and Health Policy, University of Connecticut Health Center, Farmington, CT, USA

S.H. Chiou
Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

J. Yan
Department of Statistics, University of Connecticut, Storrs, CT, USA

Center for Public Health and Health Policy, University of Connecticut Health Center, Farmington, CT, USA

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_22

1 Introduction

Every year over 200,000 people, on average, are hospitalized in the United States with complications attributed to seasonal influenza virus infections (Thompson et al. 2004). Unlike some other types of adverse health events, influenza has a seasonal pattern which tends to spike between December and February and tails off until as late as May (http://www.cdc.gov/flu/about/season/flu-season.htm). Nonetheless, abnormal outbreaks beyond the seasonal trend are not uncommon. For example, between April and October of 2009, the Center for Disease Control and Prevention (CDC) estimated between 63,000 and 153,000 people were hospitalized due to the H1N1 virus (Shrestha et al. 2011). Another example is the early outbreak of flu season in January 2013 in the New England region, when the State of Massachusetts declared a public health emergency (http://www.usatoday.com/story/news/nation/2013/01/09/boston-declares-flu-emergency/1820975/). Unexpected outbreaks result not only in more hospitalizations but also a very high concentration of hospitalizations in a very small window of time, posing challenges to health care resources and administrations. Timely detection of abnormal outbreaks of influenza is important for proactive resource allocation, disease control, and cost reduction.

Statistical surveillance aims to detect changes in a stochastic process of interest at an unknown time point as quickly and as accurately as possible. Methods of statistical surveillance originated in industrial production control, but have been applied to the context of medicine and public health (e.g., Sonesson and Bock 2003; Woodall 2006; Tennant et al. 2007; Thor et al. 2007; Coory et al. 2008; Mohammed et al. 2008). In the US, the National Notifiable Diseases Surveillance System (NNDSS) of the CDC tracks 52 different diseases on a weekly basis both at a state and a national level (Thompson et al. 2004). It has been proposed that control charts could be used on reported influenza-like illness (ILI) cases to detect outbreaks that deviate from the regular seasonal trend (Steiner et al. 2010). One limitation of the clinical and laboratory based CDC data, however, is that there may be time lags in reporting, which makes them less useful for timely detection of outbreaks. By the time the spike in hospitalizations becomes noticed, it may be too late to prepare hospitals since they may have already been flooded with patients.

A potential source for timely, inexpensive data for statistical surveillance of influenza, is the crowdsourced information from the Internet (e.g., Milinovich et al. 2014). Latest advances in information technology automate collection of higher volumes of real-time internet data that can be used as surrogate of clinically-based ILI reporting. Google Flu Trends (GFT) released counts of certain search queries that were found to be good indicators of flu activity for many regions worldwide (https://www.google.org/flutrends/about/). Ginsberg et al. (2009) used such aggregated Google search data to estimate current flu activity, measured by the proportion of physician visits attributed to influenza, around the world in near real-time; their pattern was found to match the observed CDC data very closely, with an average correlation of 0.9 across the various regions of the United States (Ginsberg et al. 2009). Similar ideas have been used with Yahoo searches (Polgreen et al. 2008), wikipedia usage (McIver and Brownstein 2014), and social media activities like twitter (Chew and Eysenbach 2010; Santos and Matos 2014). The GFT project has made a large impact in influenza forecasting (e.g., Dukic et al. 2012; Shaman and Karspeck 2012; Freyer et al. 2013; Nsoesie et al. 2013; Amorós et al. 2015).

In an alarm system, it is the abnormality detection instead of occurrence prediction that is the primary interest. Although it has been reported that models using the Google flu data overestimated influenza related clinical visit rates, and often missed deviations in the normal trend due to abnormal flu seasons (Butler 2013; Olson et al. 2013; Lazer et al. 2014), the imperfect GFT data for prediction may still be extremely useful for their real time feature in constructing an alarm system. Based on the rationale that the search queries are indicators of flu activities, an alarm system can be devised using the GFT data to warn the community of imminent abnormal influenza activity. After the initialization of GFT, the frequency data of the search terms used to predict the influenza season had been publicly available and updated regularly for public use. In August of 2015, Google ceased to publish their data, but does continue to collect and make available this data to anyone who is interested (http://googleresearch.blogspot.com/2015/08/the-next-chapter-for-flu-trends.html). The historical data for 2003 through August of 2015 are available from the GFT web page (https://www.google.org/flutrends/about/). Although data after August 2015 are not published, they can still be streamed for surveillance use with our proposed methods.

In collaboration with the Connecticut Hospital Association on public health monitoring, we develop an alarm system with statistical process control (SPC) techniques using the GFT data in Connecticut. The goal is to give an "alarm" when the GFT information suggests a deviation from the normal pattern of the influenza season. Specifically for influenza, the seasonal trend is of major interest for anticipated resource management, and it needs to be modeled with the temporal dependence appropriately accounted for. Many different control charts, each with differing capabilities, are highly applicable to health care (e.g., Faltin et al. 2012; Thor et al. 2007). The exponentially weighted moving average chart for stationary data (EWMAST) (Zhang 1998) is used here for its simplicity, with no need to model the stationary time series of the residuals from the trend model. The GFT weekly data were divided into a training set and a testing set, and a EWMAST chart was developed with the training set. The chart was applied to the testing set, and was able to issue an alarm almost immediately for the early outbreak in January 2013. The methodology can be applied to other cities or States provided that appropriate data are available—a usage of the GFT data with potentially high impact. The methods are implemented using `R`.

The rest of the paper is organized as follows. The methodologies are presented in Sect. 2, which consist of the GFT data, trend modeling and the EWMAST chart. The results are reported in detail, and validated with the testing data and CDC data in Sect. 3. A discussion concludes in Sect. 4.

2 Methods

2.1 GFT Data

Weekly counts of search queries that GFT found to be related to influenza are available at both the national and state level. Ginsberg et al. (2009) described how the search terms used by GFT were determined to be the most correlated with a region's ILI incidence rate and were selected from a pool of over fifty million possible queries. The observed CDC incidence rate was regressed onto each one of these fifty million queries to determine the 100 most individually correlated queries. From these top performers, models using the top N, $N = 1, \dots, 100$, performers were used to estimate the observed CDC data and from these models one was selected to determine the search queries best suited for tracking (Ginsberg et al. 2009). The resulting counts are our departure point.

Figure 1 shows the weekly query counts on the log scale for Connecticut from September, 2003 to August, 2015. The weekly data clearly has a seasonal pattern; the trend seems to peak in February and bottom out in June in a given year. It is also noted that there are clear anomalies in the data, especially during the H1N1 outbreak in 2009. To evaluate the performance of the EWMAST chart, data from 2013 through 2015 were left as the testing set, which contains an early peak in January 2013; the rest of the data were used as training data.

The basic concept of a control chart is to take data that is considered "in control" or representative of the normal state to develop appropriate limits such that if data were observed beyond these limits, the system would be considered out of control. Therefore, the training data needs to be divided into "in control" and "out of control" sections; only the "in control" data will be used in the construction of the EWMAST chart. A criterion is needed to determine if an observation is "in control". Ideally, some objective assessment or experts' opinion could be used. This is not as clear

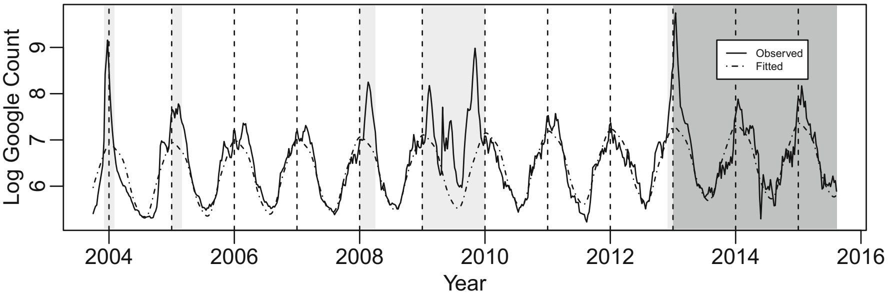


Fig. 1 Google data. This chart shows the original Google search count data on a logarithmic scale. The testing region is highlighted in *dark gray* in the chart, the rest constitutes the training regions. The training data is overlaid by the best periodic model (selected by AIC), and the regions that were deemed "out of control" are highlighted in *light gray*

cut as one might hope because if that were a straightforward task, there would be no need of any alarm system. Therefore, the data was examined retrospectively and sections of the data where we would have liked to have been notified that something was out of the ordinary are identified as “out of control”. To do this, we model the seasonal trend using the whole training data in the next subsection; observations that we would like to detect were considered as “out of control” and the rest were considered “in control”.

2.2 Modeling the Trend

Let $Y(t)$ be the log count of Google queries at time t, where t is in continuous time but $Y(t)$ is only discretely observed. Consider a model that captures a linear long-term trend and periodic seasonal trend:

$$Y(t) = \alpha_0 + \alpha_1 t + \sum_{i=1}^{k} \left[\eta_i \cos(2^i \pi t) + \beta_i \sin(2^i \pi t)\right] + \epsilon_t, \tag{1}$$

where α_0 is the intercept, α_1 is the multiplicative linear coefficient, η_i and β_i are the multiplicative periodic coefficients of degree i, k is the highest degree of periodicity considered, and $\epsilon_t \sim N(0, \sigma^2)$ is the error term.

Model (1) can be fitted to the training data with a sequence of k values, and the models can be compared with the Akaike Information Criterion (AIC) (Akaike 1974). The model with the smallest AIC will be used as a benchmark to pick out “out of control” data that one would like to detect. Then, model (1) will be fitted to the “in control” data with a sequence of k values, and the model with the optimal AIC value will be used as the trend to give a stationary residual series of the “in control” data. Stationarity can be checked with standard tests such as the Kwiatkowski–Phillips–Schmidt–Shin test (Kwiatkowski et al. 1992), augmented Dickey–Fuller test (Said and Dickey 1984), and the Phillips–Perron test (Perron 1988; Phillips and Perron 1988). Once the residuals are confirmed to be a stationary process, the EWMAST chart will be constructed based on them.

2.3 EWMAST Chart

The EWMAST chart generalizes the traditional exponentially weighted moving average (EWMA) chart for independent and identically distributed data to stationary processes (Zhang 1998). Let X_t be observations of a stationary process. In this application, it is the residual of the “in control” log count data with the fitted trend removed. The EWMAST chart does not need to model the stationary process. Define EWMAST statistic as

$$Z_t = (1 - \lambda) Z_{t-1} + \lambda X_t, \quad t = 1, 2, \ldots, \tag{2}$$

where $0 < \lambda < 1$ is the weight given to the most recent X_t. The EWMAST chart is constructed by plotting Z_t and comparing the points to the control lines established at

$$\mu \pm L\sigma_z, \tag{3}$$

where μ is the mean of $\{X_t\}$, L is a predetermined scaling factor, and σ_z^2 is the variance of $\{Z_t\}$. The mean parameter μ can be estimated with the sample mean of $\{X_t\}$, $\bar{X}$. The variance σ_z can be estimated by

$$\hat{\sigma}_z^2 = \frac{\lambda}{2-\lambda}\hat{\sigma}_x^2 \left(1 + 2\sum_{k=1}^{M} \hat{\rho}(k)(1-\lambda)^k \left[1-(1-\lambda)^{2(M-k)}\right]\right), \tag{4}$$

where $\hat{\sigma}_x^2$ is an estimate of the variance of $\{X_t\}$, $\hat{\rho}(k)$ is the sample autocorrelation of $\{X_t\}$ at lag k, and M an integer large enough such that $\hat{\sigma}_z^2$ is stable (Zhang 1998).

Zhang recommends selecting $\lambda = .2$ and M large enough make the approximation of σ_z good, while still being less than one-fourth the total data set (Zhang 1998). Since we are only interested in alarms for outbreaks, we will focus on the upper limit only. If a Z_t goes beyond the upper limit, then the point is deemed out of control, and an alarm is issued. Since the control limit is determined by the scaling factor L, for this paper we will use $L = 4$. One may then compare where the Google data indicates that there is an alarming situation to the CDC data on influenza hospitalizations at the same time to see if the use of the control chart on the Google data can identify and predict spikes in influenza hospitalizations.

3 Results

3.1 Trends

The training data needs to be split into "in control" and "out of control" portions. We did this by removing regions where we would have wanted to be notified if we were receiving the data currently, and labeled those regions as "out of control". To determine whether we would have wanted to be notified, we fit the best model to $Y(t)$ based on AIC. The approach described in Sect. 2 was used to do this and the best model (1) chosen by AIC has $k = 1$. Figure 1 shows fitted model over the original data. We determined sections of the plot that would have been alarming to us to be those where the residuals were strikingly large (those regions highlighted in Fig. 1). These sections were treated as "out of control" while the remaining portion of the data was designated as "in control" and was then used to develop EWMAST chart.

Model 1 was fitted to the "in control" data with $k \in \{1, \ldots, 5\}$, and the AIC values for each model are summarized Table 1. The best model with the smallest AIC has $k = 3$. The coefficient estimates and their standard errors are summarized

Table 1 AIC values for models with varying degrees of periodicity considered

Model	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
AIC	23.48	15.48	15.01	17.33	21.26

Table 2 Coefficient estimates and standard errors for the trend model

Term	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\eta}_1$	$\hat{\beta}_1$
Estimate(SE)	−76.88 (8.88)	0.04 (0.00)	0.76 (0.02)	0.22 (0.02)
Term	$\hat{\eta}_2$	$\hat{\beta}_2$	$\hat{\eta}_3$	$\hat{\beta}_3$
Estimate(SE)	−0.04 (0.02)	−0.04 (0.02)	0.04 (0.02)	−0.01 (0.02)

in Table 2. All coefficients but the third order periodic terms ($\hat{\eta}_3$ and $\hat{\beta}_3$) in Table 2 are found to be statistically significant at the 0.05 significance level. We note the significance of the linear term especially because it indicates that there is a slight increase in influenza related search terms over the years from 2003 until 2012. This is likely due to a general increase in the use of the Internet over these years, but it remains noteworthy.

Before using the residuals from the optimal model to develop the EWMAST chart, the residuals must be ensured to be a stationary process. The Kwiatkowski–Phillips–Schmidt–Shin Test tests the null hypothesis that the process is stationary, while the Phillips–Perron test and the augmented Dickey–Fuller test both test the null hypothesis that the process has a unit root, which means that the process is not stationary. The Kwiatkowski–Phillips–Schmidt–Shin test fails to reject with a p-value $P > 0.1$ while the augmented Dickey–Fuller test and Phillips–Perron test do reject with a p-value $P < 0.01$. Thus, it is concluded that the residuals can in fact be considered a stationary process.

3.2 *EWMAST Chart*

The control limit depends on estimates of μ and σ_z of the transformed series Z_t in Eq. (2). With $\lambda = .2$ suggested by Zhang (1998), the residuals after removing the estimated trend in (1) were used as input X_t to obtain the transformed series Z_t; see Fig. 3, Panel (a). The mean parameter μ was estimated as $\hat{\mu} = 0$. To select M in estimation of σ_z^2 in (4), the estimate $\hat{\sigma}_z^2$ based on different M values is plotted against $M \in \{1, \ldots, 30\}$, in Fig. 2. The estimate $\hat{\sigma}_z^2$ stabilized at $M = 20$. With $L = 4$, the upper control limit is determined to be 0.53.

Normally, the control limit L would be tuned based on a control chart performance measure such as average run length (Zhang 2000; Han and Tsung 2009) and sample size (Köksal et al. 2008; Capizzi and Masarotto 2007) with potential adjustments to account for estimation error (Apley and Cheol Lee 2003). Because of the nature by which our chart is constructed, none of these approaches can be utilized in a meaningful way. Instead we opt to set $L = 4$ as a very conservative bound that even under Chebyshev's inequality would give the chance of a false alarm a lower bound of 6.25 %.

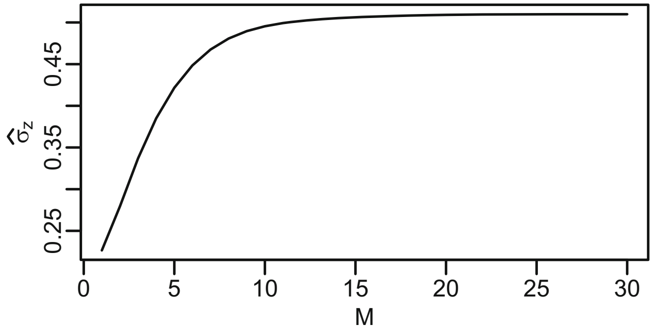


Fig. 2 Estimates of EWMAST standard deviation for increasing lag times

The EWMAST chart along with the original GFT data with "out of control" regions highlighted are presented the first two panels of Fig. 3. For the training data, the EWMAST chart issues alarm for "out of control" observations in a timely manner. The regions where the control chart alarmed as out of control match closely with the timings of the "out of control" data. For example, the chart was able to detect the H1N1 outbreak in 2009.

3.3 Validation

The EWMAST chart constructed and tested with the GFT data still needs to be validated with the real influenza incidences. When the alarm system alerts will be compared to the incidence rates in Connecticut as reported by the CDC during those times. Panel (c) of Fig. 3 presents the weekly incident rates per 100 thousands population of laboratory confirmed influenza hospitalization in Connecticut for the flu seasons of 2004–2015 from the CDC. The CDC only collected these data during the traditional flu season from the beginning of October to the end of March. The EWMAST chart is also presented for side-by-side comparison. To evaluate this system, where the alarm system correctly identifies points in time when there is an unusual event, such as a higher than normal hospitalization rate (a spike), or a shift in the normal flu season is examined. Where the system incorrectly identifies abnormal situations is also considered.

It is seen that the primary events to be concerned with are the spikes just before 2004, 2010 (H1N1), and right at the beginning of 2013. These regions are of interest both for their shift in time and for the sheer number of hospitalizations during the peak, as compared with more typical seasons.

The system does a fairly good job identifying the spike at the beginning of 2013. The control limit was exceeded in late December, 2012, and the monitoring series Z_t stayed outside the limit for most part of the flu season in 2013. This early season, a likely extension of the Massachusetts public health emergency, would have been detected early by this system, which in turn would have allowed for more time to prepare.

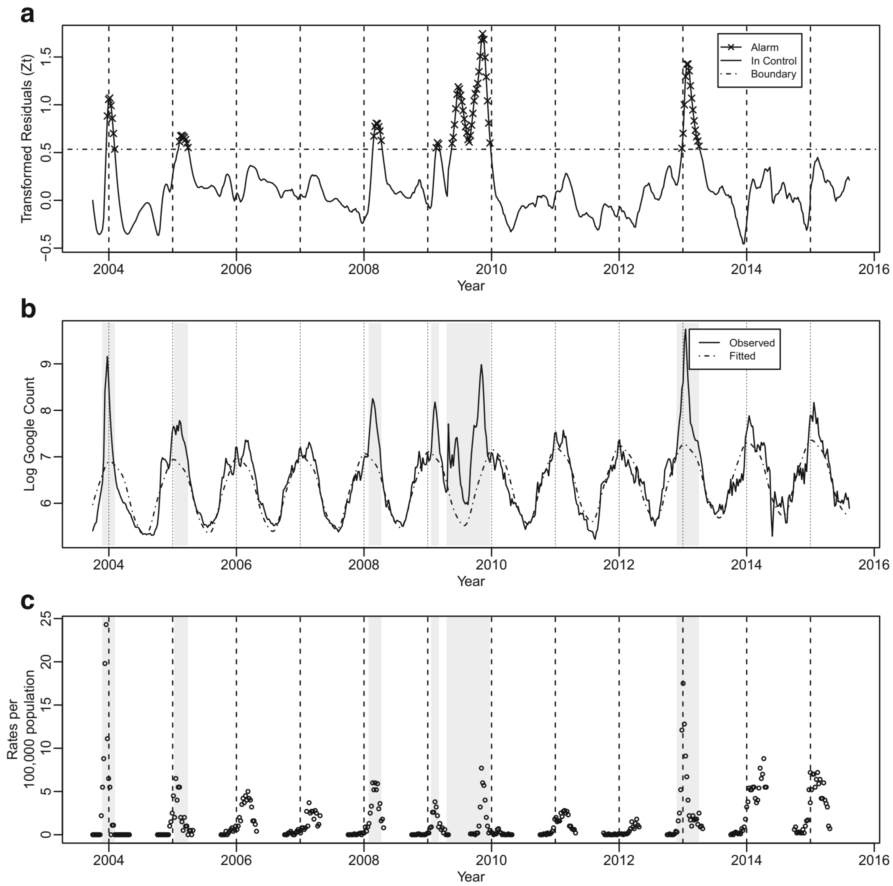


Fig. 3 EWMAST chart in comparison with the observed GFT data and the CDC laboratory confirmed influenza hospitalizations per 100,000 people in the State of Connecticut. Panel **a** is the EWMAST chart of the Z_t series, the *horizontal dot-dash line* representing the upper control limit. The Z_t series constructed from the seasonal model's residuals are plotted and the "×" markers are overlaid to indicate a data point over the control limit, and therefore "out of control". Panel **b** shows the GFT count data, with "out of control" data highlighted in *gray*. Panel **c** presents the weekly number of laboratory confirmed influenza hospitalizations per 100,000 people in the population for the state of Connecticut (data from the CDC), with regions that the EWMAST chart identified as out of control highlighted in *gray*

The system also does a great job identifying the spike before 2010 (the H1N1 outbreak), again covering the time of the spike as well as the time before it. It is noted that the system alarms well before the spike, which is not as helpful, but is to be expected, given that the outbreak began earlier in 2009 in Mexico, and there was a great deal of anticipation of an outbreak in the U.S. The system does alarm during the 2003–2004 spike, but it only signals as higher than normal during the actual peak.

The system also identifies the peaks in 2005 and 2008. While the peak in 2005 did occur closer to when the peak for the season was expected, the alarm correctly indicates that the peak in 2005 is slightly early. The alarm in 2008 however is somewhat unneeded as the spike is neither particularly large nor particularly early.

In the remaining years of the testing data, we see that there are no false alarms. While the peak in 2015 does arrive a little early, neither peak in 2014 nor in 2015 is particularly large, and thus the lack of alarm is not a detriment.

4 Discussion

Seasonal influenza poses as a serious concern to public health. The seasonal trend can be estimated, but unpredictable factors are always at play that can alter the timing and severity of the peak. We present a simple, quick, and inexpensive approach that, instead of trying to estimate the current severity of the influenza season like previous techniques, simply anticipates abnormal occurrences in the influenza season by making use of the timely, inexpensive GFT data. Such a source does not suffer from the delay of the other primary source of influenza reporting which may be more accurate but slow. The EWMAST control chart can alarm public health officials when an unexpected influx of ILI cases may be looming. This approach could serve to inform hospitals in time to be better prepared for abnormality from the expected seasonal pattern.

While this paper serves as a good first step, further research is merited to improve upon what has been presented here. The process by which the training data was determined was noticeably subjective; an more objective method would be better in selecting the training data and determination of “in control”. It may be worth reproducing the case study in other regions using GFT data or data from other search engines or social media. As time goes on, the control limit can be adjusted by repeating the process described with additional most recent data as necessary to reduce the effect of parameter estimation uncertainty and to allow for update in possible changes in trend and seasonal pattern. The presented technique could also be used directly or in conjunction with hospitalization counts to help further improve the detection and prediction of spikes of influenza or potentially other seasonal disease hospitalizations, again with the goal of giving as early warning as possible to hospitals.

Conflict of Interest The authors have declared no conflict of interest.

References

Akaike, H. (1974). A New Look at the Statistical Model Identification. *Automatic Control, IEEE Transactions on* **19**, 716–723.

Amorós, R., Conesa, D., Martinez-Beneito, M. A., and López-Quılez, A. (2015). Statistical Methods for Detecting the Onset of Influenza Outbreaks: A Review. *REVSTAT—Statistical Journal* **13**, 41–62.

Apley, D. W. and Cheol Lee, H. (2003). Design of Exponentially Weighted Moving Average Control Charts for Autocorrelated Processes with Model Uncertainty. *Technometrics* **45**, 187–198.

Butler, D. (2013). When Google Got Flu Wrong. *Nature* **494**, 155.

Capizzi, G. and Masarotto, G. (2007). The EWMAST Control Charts with Estimated Limits: Properties and Recommendations. In *Industrial Engineering and Engineering Management, 2007 IEEE International Conference on*. IEEE, pages 1403–1407.

Chew, C. and Eysenbach, G. (2010). Pandemics in the Age of Twitter: Content Analysis of Tweets During the 2009 H1N1 Outbreak. *PloS one* **5**, e14118.

Coory, M., Duckett, S., and Sketcher-Baker, K. (2008). Using Control Charts to Monitor Quality of Hospital Care with Administrative Data. *International Journal for Quality in Health Care* **20**, 31–39.

Dukic, V., Lopes, H. F., and Polson, N. G. (2012). Tracking Epidemics with Google Flu Trends Data and A State-Space SEIR Model. *Journal of the American Statistical Association* **107**, 1410–1426.

Faltin, F., Kenett, R., and Ruggeri, F. (2012). *Statistical Methods in Healthcare*. John Wiley & Sons, New York.

Freyer, A., Jalalpour, M., Gel, Y., Levin, S., and Torcaso, F. (2013). Influenza Forecasting with Google Flu Trends. *PloS one* **8**, e56176.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., *et al.* (2009). Detecting Influenza Epidemics Using Search Engine Query Data. *Nature* **457**, 1012–1014.

Han, D. and Tsung, F. (2009). Run Length Properties of the CUSUM and EWMA Schemes for a Stationary Linear Process. *Statistica Sinica* **19**, 473.

Köksal, G., Kantar, B., Ali Ula, T., and Caner Testik, M. (2008). The Effect of Phase I Sample Size on the Run Length Performance of Control Charts for Autocorrelated Data. *Journal of Applied Statistics* **35**, 67–87.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure are We That Economic Time Series Have a Unit Root? *Journal of Econometrics* **54**, 159–178.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science* **343**, 1203–1205.

McIver, D. J. and Brownstein, J. S. (2014). Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. *PLoS Computational Biology* **10**, e1003581.

Milinovich, G. J., Williams, G. M., Clements, A. C. A., and Hu, W. (2014). Internet-Based Surveillance Systems for Monitoring Emerging Infectious Diseases. *The Lancet Infectious Diseases* **14**, 160–168.

Mohammed, M., Worthington, P., and Woodall, W. (2008). Plotting Basic Control Charts: Tutorial Notes for Healthcare Practitioners. *Quality and Safety in Health Care* **17**, 137–145.

Nsoesie, E., Mararthe, M., and Brownstein, J. (2013). Forecasting Peaks of Seasonal Influenza Epidemics. *PLoS Currents* **5**.

Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., and Simonsen, L. (2013). Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. *NVM* **9**, e1003256.

Perron, P. (1988). Trends and Random Walks in Macroeconomic Time Series: Further Evidence From a New Approach. *Journal of Economic Dynamics and Control* **12**, 297–332.

Phillips, P. C. and Perron, P. (1988). Testing for a Unit Root in Time Series Regression. *Biometrika* **75**, 335–346.

Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., and Weinstein, R. A. (2008). Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases* **47**, 1443–1448.

Said, S. E. and Dickey, D. A. (1984). Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika* **71**, 599–607.

Santos, J. C. and Matos, S. (2014). Analysing Twitter and Web Queries for Flu Trend Prediction. *Theoretical Biology and Medical Modelling* **11**, S6.

Shaman, J. and Karspeck, A. (2012). Forecasting Seasonal Outbreaks of Influenza. *Proceedings of the National Academy of Sciences* **109**, 20425–20430.

Shrestha, S. S., Swerdlow, D. L., Borse, R. H., Prabhu, V. S., Finelli, L., *et al.* (2011). Estimating the Burden of 2009 Pandemic Influenza A (H1N1) in the United States (April 2009–April 2010). *Clinical Infectious Diseases* **52**, S75–S82.

Sonesson, C. and Bock, D. (2003). A Review and Discussion of Prospective Statistical Surveillance in Public Health. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **166**, 5–21.

Steiner, S. H., Grant, K., Coory, M., and Kelly, H. A. (2010). Detecting the Start of an Influenza Outbreak Using Exponentially Weighted Moving Average Charts. *BMC Medical Informatics and Decision Making* **10**, 37.

Tennant, R., Mohammed, M. A., Coleman, J. J., and Martin, U. (2007). Monitoring Patients Using Control Charts: A Systematic Review. *International Journal for Quality in Health Care* **19**, 187–194.

Thompson, W. W., Shay, D. K., Weintraub, E., Brammer, L., Bridges, C. B., *et al.* (2004). Influenza-Associated Hospitalizations in the United States. *JAMA* **292**, 1333–1340.

Thor, J., Lundberg, J., Ask, J., Olsson, J., Carli, C., *et al.* (2007). Application of Statistical Process Control in Healthcare Improvement: Systematic Review. *Quality and Safety in Health Care* **16**, 387–399.

Woodall, W. H. (2006). The Use of Control Charts in Health-Care and Public-Health Surveillance. *Journal of Quality Technology* **38**, 89–104.

Zhang, N. F. (1998). A Statistical Control Chart for Stationary Process Data. *Technometrics* **40**, 24–38.

Zhang, N. F. (2000). Statistical Control Charts for Monitoring the Mean of a Stationary Process. *Journal of Statistical Computation and Simulation* **66**, 249–258.

Identifying Gene-Environment Interactions with a Least Relative Error Approach

Yangguang Zang, Yinjun Zhao, Qingzhao Zhang, Hao Chai, Sanguo Zhang, and Shuangge Ma

Abstract For complex diseases, the interactions between genetic and environmental risk factors can have important implications beyond the main effects. Many of the existing interaction analyses conduct marginal analysis and cannot accommodate the joint effects of multiple main effects and interactions. In this study, we conduct joint analysis which can simultaneously accommodate a large number of effects. Significantly different from the existing studies, we adopt loss functions based on relative errors, which offer a useful alternative to the "classic" methods such as the least squares and least absolute deviation. Further to accommodate censoring in the response variable, we adopt a weighted approach. Penalization is used for identification and regularized estimation. Computationally, we develop an effective algorithm which combines the majorize-minimization and coordinate descent. Simulation shows that the proposed approach has satisfactory performance. We also analyze lung cancer prognosis data with gene expression measurements.

Keywords Gene-environment interactions • Robustness • Least relative error • Cancer

1 Introduction

For complex diseases, it is of significant interest to identify genetic risk factors. For etiology, biomarkers, and prognosis, the interactions between genetic and environmental risk factors, also referred to as G × E interactions, have important implications beyond the main effects. Extensive studies have been conducted to search for important G × E interactions (Cordell 2009; Hunter 2005; North and

Y. Zang • S. Zhang
School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China
e-mail: zangyangguang@mails.ucas.ac.cn; sgzhang@ucas.ac.cn

Y. Zhao • Q. Zhang • H. Chai • S. Ma (✉)
Department of Biostatistics, Yale University, New Haven, CT, USA
e-mail: yinjun.zhao@yale.edu; qzzhang@ucas.ac.cn; hao.chai@yale.edu; shuangge.ma@yale.edu

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_23

Martin 2008; Wu and Ma 2015). In this article, we focus on analyzing prognosis, which has an essential role in biomedical studies. It is conjectured that the proposed approach can be extended to some types of disease outcomes.

Denote T as the survival time, $Z = (Z_1, \cdots, Z_p)^\top \in R^{p\times 1}$ as the p genetic factors (G), and $X = (X_1, \cdots, X_q)^\top \in R^{q\times 1}$ as the q clinical/environmental risk factors (E). There are two families of methods for detecting the main G and E effects and G×E interactions. The first conducts marginal analysis and analyzes one or a small number of G factors at a time (Hunter 2005; Shi et al. 2014; Thomas 2010). With a slight abuse of terminology, we use the generic phrase "gene" for the G factor. In marginal analysis, for gene k, consider the model $T \sim \phi(\sum_{j=1}^{q} X_j\alpha_j + Z_k\beta_k + \sum_{j=1}^{q} X_jZ_k\xi_{jk})$, where model ϕ is known up to the regression coefficients, and $\alpha_j, \beta_k, \xi_{jk}$ are the unknown regression coefficients. Marginal analysis is easy to implement however cannot accommodate the joint effects of multiple main effects and interactions. The second family conducts joint analysis and describes the joint effects of all factors in a single model (Liu et al. 2013; Wu et al. 2014; Zhu et al. 2014). More specifically, consider $T \sim \phi(\sum_{j=1}^{q} X_j\alpha_j + \sum_{k=1}^{p} Z_k\beta_k + \sum_{j=1}^{q}\sum_{k=1}^{p} X_jZ_k\xi_{jk})$.

For the simplicity of description, consider the simple linear regression setting. In the literature, most of the existing methods adopt loss functions built on *absolute error criteria*, with the most popular including the least squares (LS) and least absolute deviation (LAD). Under certain settings, it has been found that the *relative errors* are more sensible (Khoshgoftaar et al. 1992; Park and Stefanski 1998; Van Dam and Ernst 2015). The most distinguishable feature of the relative error-based approaches is that they are scale-free, which, as discussed in the published studies (Chen et al. 2010), can be advantageous in survival and other analysis. There are at least two ways of defining relative error-based criteria. The first is defined based on the ratio of the error with respect to the target. The second is defined on the ratio of the error with respect to the predictor (Chen et al. 2010). Based on the two types of relative errors, researchers have proposed the least absolute relative errors (LARE) criterion and the least product relative errors (LPRE) criterion for linear multiplicative models. The LARE criterion is convex but not smooth. For its extensions and applications, we refer to Li et al. (2014), Tsionas (2014), and Zhang and Wang (2013) and followup studies. In comparison, the LPRE criterion is smooth and convex (Chen et al. 2013). Under low-dimensional settings, asymptotical properties of the LARE and LPRE estimates for linear multiplicative models have been established (Chen et al. 2010, 2013).

Different from the existing ones (which focus on the main effects), this study adopts the relative error-based criteria for analyzing interactions. Such new criteria may provide a useful alternative to the commonly-adopted absolute error-based criteria. In genetic data analysis, it is critical to identify the important main effects and interactions, which poses a variable (model) selection problem. In two recent studies (Xia et al. 2015; Zhang and Wang 2013), variable selection based on the LARE has been studied. However, the existing studies are limited to the situation where the dimension of model is smaller than the sample size. To the best of our knowledge, there is a lack of study examining the relative error-based criteria under high-dimensional settings. Also different from the existing studies, we analyze prognosis data under right censoring, which introduces additional complexity.

2 Methods

2.1 Model and Relative Error-Based Criteria

For modeling a prognosis response, we consider the following linear multiplicative (accelerated failure time—AFT) model,

$$T = \exp\Big(\sum_{j=1}^{q} X_j\alpha_j + \sum_{k=1}^{p} Z_k\beta_k + \sum_{j=1}^{q}\sum_{k=1}^{p} X_jZ_k\xi_{jk}\Big)\varepsilon, \tag{1}$$

where ε is the random error independent of X and Z. This model provides a useful alternative to the Cox and other models. It can be especially preferred under high-dimensional settings. Let $U = (X^\top, Z^\top, (X \otimes Z)^\top)^\top$ and $\theta = (\alpha^\top, \beta^\top, \xi^\top)^\top$, then we can write model (1) as

$$T = \exp(U^\top\theta)\varepsilon. \tag{2}$$

First consider the case without censoring. Suppose that we have n iid observations $\{t_i, \mathbf{x}_i, \mathbf{z}_i\}_{i=1}^n$, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{iq})^\top$ and $\mathbf{z}_i = (z_{i1}, \cdots, z_{ip})^\top$. Denote $\mathbf{u}_i = (\mathbf{x}_i^\top, \mathbf{z}_i^\top, (\mathbf{x}_i \otimes \mathbf{z}_i)^\top)^\top$. With the logarithm transformation, model (2) can be rewritten as $\log(T) = U^\top\theta + \log(\varepsilon)$. The LS and LAD methods can be applied, which, respectively, minimize the objective functions $\sum_{i=1}^n(\log(t_i) - \mathbf{u}_i^\top\theta)^2$ and $\sum_{i=1}^n |\log(t_i) - \mathbf{u}_i^\top\theta|$. Both methods are built on the absolute errors.

As discussed in the literature, under certain scenarios, the relative error-based criteria can be more sensible. In this article, we consider the least absolute relative errors (LARE) (Chen et al. 2010) and least product relative errors (LPRE) (Chen et al. 2013) criteria. They have been relatively more popular in the relative error literature and deserve a higher priority. The LARE objective function is defined as

$$LARE_n(\theta) = \sum_{n=1}^{n}\left\{\left|\frac{t_i - \exp(\mathbf{u}_i^\top\theta)}{t_i}\right| + \left|\frac{t_i - \exp(\mathbf{u}_i^\top\theta)}{\exp(\mathbf{u}_i^\top\theta)}\right|\right\}. \tag{3}$$

The LPRE objective function is defined as

$$LPRE_n(\theta) = \sum_{n=1}^{n}\left\{\left|\frac{t_i - \exp(\mathbf{u}_i^\top\theta)}{t_i}\right| \times \left|\frac{t_i - \exp(\mathbf{u}_i^\top\theta)}{\exp(\mathbf{u}_i^\top\theta)}\right|\right\}. \tag{4}$$

Now consider the realistic case with right censoring. For subject $i(= 1, \ldots, n)$, let c_i be the censoring variable which is independent of $\mathbf{x}_i, \mathbf{z}_i$, and t_i. We observe $y_i = \min(t_i, c_i)$ and $\delta_i = 1(t_i \le c_i)$. Without loss of generality, assume that the data $(y_i, \delta_i, \mathbf{u}_i)$ have been sorted according to y_i from the smallest to the largest.

2.2 Penalized Estimation and Selection

Consider the general relative error (GRE) criterion

$$GRE_n(\theta) = \sum_{n=1}^{n} g\left\{\left|\frac{t_i - \exp(\mathbf{u}_i^\top\theta)}{t_i}\right|, \left|\frac{t_i - \exp(\mathbf{u}_i^\top\theta)}{\exp(\mathbf{u}_i^\top\theta)}\right|\right\}, \tag{5}$$

where $g(a, b)$ is a bivariate function satisfying certain regularity conditions. When $g(a, b) = a + b$, the GRE criterion becomes the LARE (Chen et al. 2010); when $g(a, b) = ab$, it becomes the LPRE (Chen et al. 2013).

To accommodate right censoring in estimation, we adopt a weighted approach. Specifically, we first compute the Kaplan-Meier weights $\{w_i\}_{i=1}^n$ as

$$w_1 = \frac{\delta_1}{n}, w_i = \frac{\delta_i}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_j}, i = 2, \cdots, n. \tag{6}$$

We propose the weighted objective function

$$Q_n(\theta) = \sum_{i=1}^{n} w_i g\left\{\left|\frac{y_i - \exp(\mathbf{u}_i^\top\theta)}{y_i}\right|, \left|\frac{y_i - \exp(\mathbf{u}_i^\top\theta)}{\exp(\mathbf{u}_i^\top\theta)}\right|\right\}. \tag{7}$$

In genetic interaction analysis, the dimension of unknown parameters can be much larger than the sample size. For regularized estimation and identification of important effects, we adopt penalization, where the objective function is

$$L_{n,\lambda}(\theta) = Q_n(\theta) + \varphi_\lambda(\theta). \tag{8}$$

Here $\varphi_\lambda(\theta)$ is the penalty function. Adopting penalization for genetic interaction analysis has been pursued in recent literature. See for example Bien et al. (2013), Liu et al. (2013), and Shi et al. (2014).

Multiple penalties are potentially applicable. Here we adopt the MCP (Zhang 2010), which has been the choice of many high-dimensional studies including genetic interaction analysis. The penalty is defined as $\varphi_\lambda(t) = \lambda\int_0^{|t|}(1 - x/(\gamma\lambda))_+ dx$. $\gamma > 0$ is the regularization parameter, and λ is the tuning parameter.

It is noted that applying the MCP may lead to results not respecting the "main effects, interactions" hierarchy, which has been stressed in some recent studies (Bien et al. 2013). The hierarchy postulates that the main effects corresponding to the identified interactions should be automatically identified. This can be achieved by replacing the MCP with for example sparse group penalties. However, we note that the computational cost of such penalties can be much higher. In addition, some published studies have demonstrated pure interactions without the presence of main effects (Caspi and Moffitt 2006; Zimmermann et al. 2011). In data analysis, when it is necessary to reinforce the hierarchy, we can refit and add back the main effects corresponding to the identified interactions (if these main effects are not identified in the first place).

2.3 Computation

For optimizing the penalized objective function, we propose combining the majorize-minimization (MM) algorithm (Hunter and Li 2005) with the coordinate descent (CD) algorithm (Wu and Lange 2008). The MM is used to approximate the objective function using its quadratic majorizer, while the CD is used for iteratively updating the estimate.

Specifically, when $g(a, b) = ab$, it is easy to compute the gradient and hessian matrix for $Q_n(\theta)$, and so approximation may not be needed. However when $g(a, b) = a + b$, computing the hessian matrix becomes difficult. With the estimate $\theta^{(s)}$ at the beginning of the $s + 1$th iteration, we approximate $Q_n(\theta)$ by

$$Q_n(\theta;\theta^{(s)}) = \frac{1}{2}\sum_{i=1}^{n} w_i \left\{ \frac{(1 - y_i^{-1}\exp(\mathbf{u}_i^\top\theta))^2}{|1 - y_i^{-1}\exp(\mathbf{u}_i^\top\theta^{(s)})|} + |1 - y_i^{-1}\exp(\mathbf{u}_i^\top\theta^{(s)})| \right.$$
$$\left. + \frac{(1 - y_i\exp(-\mathbf{u}_i^\top\theta))^2}{|1 - y_i\exp(-\mathbf{u}_i^\top\theta^{(s)})|} + |1 - y_i\exp(-\mathbf{u}_i^\top\theta^{(s)})| \right\}.$$

It can be shown that $Q_n(\theta;\theta^{(s)}) \geq Q_n(\theta)$, and the equality holds if and only if $\theta^{(s)} = \theta$. For the MCP, we use a quadratic approximation

$$\varphi_\lambda(\theta;\theta^{(s)}) = \varphi_\lambda(\theta^{(s)}) + \frac{1}{2|\theta^{(s)}|}\varphi_\lambda'(\theta^{(s)})(\theta^2 - \theta^{(s)2}).$$

By ignoring terms not related to θ in $Q_n(\theta;\theta^{(s)}) + \varphi_\lambda(\theta;\theta^{(s)})$, we have a smooth loss function $L_{n,\lambda}(\theta;\theta^{(s)})$, which is

$$\sum_{i=1}^{n} w_i \left\{ \frac{(1 - y_i^{-1}\exp(\mathbf{u}_i^\top\theta))^2}{|1 - y_i^{-1}\exp(\mathbf{u}_i^\top\theta^{(s)})|} + \frac{(1 - y_i\exp(-\mathbf{u}_i^\top\theta))^2}{|1 - y_i\exp(-\mathbf{u}_i^\top\theta^{(s)})|} \right\} + \frac{1}{|\theta^{(s)}|}\varphi_\lambda'(\theta^{(s)})\theta^2. \quad (9)$$

To solve the minimization problem $\theta^{(s+1)} = \arg\min_\theta L_{n,\lambda}(\theta;\theta^{(s)})$, we employ the coordinate descent algorithm. In summary, the algorithm proceeds as follows:

Step 1. Initialize $s = 0$. Compute $\theta^{(0)}$ as the Lasso estimate (which can be viewed as an extreme case of the MCP estimate).

Step 2. Apply the CD algorithm to minimize the loss function $L_{n,\lambda}(\theta;\theta^{(s)})$ in (9). Denote the estimate as $\theta^{(s+1)}$. Specially, the CD algorithm updates one coordinate at a time and treats the other coordinates as fixed. Define u_{ij} as the jth component of $\mathbf{u}_i$. For $j \in \{1, \cdots, p+q+pq\}$, defined $\vartheta_{i,-j} = \sum_{t<j} u_{it}\theta_t^{(s+1)} + \sum_{t>j} u_{it}\theta_t^{(s)}$, then

$$\theta_j^{(s+1)} = \arg\min_{\theta_j} \left\{ \sum_{i=1}^{n} w_i \left[\frac{(1 - y_i^{-1}\exp(\vartheta_{i,-j} + u_{ij}\theta_j))^2}{|1 - y_i^{-1}\exp(\mathbf{u}_i^\top \theta^{(s)})|} + \frac{(1 - y_i \exp(-\vartheta_{i,-j} - u_{ij}\theta_j))^2}{|1 - y_i \exp(-\mathbf{u}_i^\top \theta^{(s)})|} \right] + \frac{1}{|\theta_j^{(s)}|}\varphi_\lambda'(\theta_j^{(s)})\theta_j^2 \right\} .$$

Step 3. Repeat Step 2 until convergence. We use the L_2-norm of the difference between two consecutive estimates less than 10^{-6} as the convergence criterion.

The proposed method involves tunings. For γ, published studies (Zhang 2010) suggest selecting from a small number of values or fixing it. In our simulation, we find that the estimation results are not sensitive to the value of γ. We follow published studies and set $\gamma = 6$. The selection of λ will be described in the following sections.

3 Simulation

Beyond evaluating performance of the proposed approach, we also use simulation to compare with the penalized weighted least squares (simply denoted as LS) and penalized weighted least absolute deviation (denoted as LAD) methods, which respectively have objective functions

$$\sum_{i=1}^{n} w_i(\log(y_i) - \mathbf{u}_i^\top\theta)^2 + \varphi_\lambda(\theta) \text{ and } \sum_{i=1}^{n} w_i|\log(y_i) - \mathbf{u}_i^\top\theta| + \varphi_\lambda(\theta),$$

where $\{w_i\}_{i=1}^n$ and $\varphi_\lambda(\theta)$ are the same as defined before.

Simulation I. In model $t_i = \exp(\mathbf{x}_i^\top\alpha + \mathbf{z}_i^\top\beta + (\mathbf{x}_i \otimes \mathbf{z}_i)^\top\xi)\varepsilon_i$, $i = 1, \cdots, n$, $\mathbf{z}_i$'s have a multivariate normal distribution with marginal means 0 and marginal variances 1. Denote the correlation coefficient between genes j and k as ρ_{jk}. Consider the following correlation structures: (i) independent, where $\rho_{jk} = 0$ if $j \neq k$, (ii) AR(0.2), where $\rho_{jk} = 0.2^{|j-k|}$; (iii) AR(0.8), where $\rho_{jk} = 0.8^{|j-k|}$; (iv) Band1, where $\rho_{jk} = 0.3$ if $|j-k| = 1$ and $\rho_{jk} = 0$ otherwise; and (v) Band2, where $\rho_{jk} = 0.6$ if $|j-k| = 1$, $\rho_{jk} = 0.3$ if $|j-k| = 2$, and $\rho_{jk} = 0$ otherwise. We generate $\mathbf{x}_i$'s from the standard multivariate normal distribution. We set $n = 200$, $q = 5$, and $p = 500$. The dimension of genetic effects and interactions is much larger than the sample size. There are a total of 35 nonzero effects: 5 main effects of the E factors, 10 main effects of the G factors, and 20 interactions. The nonzero coefficients are randomly generated from *Uniform*$(0.4, 1.2)$. We consider two error distributions:

(i) $\log(\varepsilon)$ follows $N(0, 1)$, and (ii) $\log(\varepsilon)$ follows $Unif(-2, 2)$. The event times are computed from the AFT model. The censoring times are generated from a uniform distribution, with a censoring rate about 20%.

Simulation II. Data are first generated in the same manner as under Simulation I. To mimic discrete genetic data (for example SNPs), we dichotomize the simulated genetic data at -1 and 0.5 to create three levels.

We evaluate the simulation results in two ways. First, we consider a sequence of λ values, evaluate identification performance at each value, and then compute the overall AUC (area under the ROC—receiver operating characteristic—curve). In addition, we also select the optimal λ using a cross validation approach and then compute the estimation squared error (SE), true positive rate (TPR), and false positive rate (TPR) at the optimal tuning. The summary based on 200 replicates is provided in Tables 1 and 3 (Appendix), respectively. Simulation suggests that, when evaluated using AUC, the four methods have similar performance. Under Simulation I, the performance is also similar in terms of SE, TPR, and FPR. However, under Simulation II, the proposed LARE and LPRE can have better performance. In addition, it is also observed that LARE may outperform LPRE, at the cost of slightly higher computer time. Overall simulation suggests that the proposed approach, especially LARE, performs comparable to or better than the alternatives. Thus it provides a “safe” choice for practical data analysis.

4 Analysis of Lung Cancer Prognosis Data

Lung cancer is the leading cause of cancer death worldwide. Genetic profiling studies have been extensively conducting, searching for genetic risk factors associated with lung cancer prognosis. Here we analyze the TCGA (The Cancer Genome Atlas) data on the prognosis of lung adenocarcinoma. The TCGA data were recently collected and published by NCI and have a high quality. The prognosis outcome of interest is overall survival. The dataset contains records on 468 patients, among whom 117 died during follow-up. The median follow-up time is 8 months.

Four E factors are included in analysis: age, gender, smoking pack years, and smoking history. All four have been suggested as associated with lung cancer prognosis in the literature. Among them, age and smoking pack years are continuous and normalized prior to analysis. Gender and smoking history are binary. A total of 436 subjects have complete E measurements. Among them, 110 died during follow-up, and the median follow-up time is 23 months. For the 326 censored subjects, the median follow-up time is 6 months.

Measurements on 18,897 gene expressions are available. To improve stability and reduce computational cost, we conduct marginal prescreening based on genes' univariate regression significance (p-value less than or equal to 0.1) and interquartile range (above the median of all interquartile ranges). Similar procedures have been

Table 1 Summary of Simulation II

		AUC	SE	TPR	FPR
$\log(\varepsilon) \sim N(0, 1)$					
Independent	LARE	0.846(0.031)	19.53(3.321)	0.601(0.063)	0.098(0.013)
	LPRE	0.837(0.032)	19.47(3.130)	0.572(0.171)	0.095(0.134)
	LAD	0.833(0.029)	20.26(3.118)	0.564(0.117)	0.084(0.026)
	LS	0.854(0.020)	20.78(2.641)	0.562(0.109)	0.076(0.013)
AR(0.2)	LARE	0.868(0.034)	17.55(3.252)	0.739(0.082)	0.103(0.018)
	LPRE	0.863(0.024)	16.68(3.671)	0.649(0.153)	0.062(0.027)
	LAD	0.847(0.027)	19.57(2.947)	0.564(0.100)	0.078(0.026)
	LS	0.860(0.024)	18.66(2.583)	0.628(0.086)	0.071(0.011)
AR(0.8)	LARE	0.928(0.029)	7.655(2.611)	0.891(0.053)	0.062(0.027)
	LPRE	0.898(0.032)	7.755(2.990)	0.871(0.076)	0.066(0.021)
	LAD	0.911(0.022)	13.68(2.973)	0.758(0.098)	0.069(0.023)
	LS	0.901(0.026)	12.74(2.417)	0.779(0.104)	0.063(0.019)
Band1	LARE	0.868(0.033)	18.51(3.316)	0.673(0.080)	0.078(0.022)
	LPRE	0.859(0.026)	17.78(3.560)	0.641(0.143)	0.059(0.023)
	LAD	0.850(0.031)	19.27(3.676)	0.629(0.119)	0.085(0.025)
	LS	0.864(0.022)	18.92(2.853)	0.616(0.074)	0.078(0.012)
Band2	LARE	0.904(0.028)	10.82(2.571)	0.828(0.158)	0.060(0.017)
	LPRE	0.875(0.031)	11.39(2.922)	0.787(0.102)	0.055(0.021)
	LAD	0.872(0.033)	17.68(3.673)	0.685(0.108)	0.075(0.027)
	LS	0.880(0.025)	16.92(3.114)	0.725(0.081)	0.075(0.014)
$\log(\varepsilon) \sim Unif(-2, 2)$					
Independent	LARE	0.840(0.032)	19.38(3.024)	0.634(0.073)	0.111(0.024)
	LPRE	0.845(0.022)	20.46(2.898)	0.582(0.169)	0.094(0.035)
	LAD	0.831(0.033)	21.29(3.453)	0.569(0.123)	0.081(0.027)
	LS	0.847(0.021)	21.03(3.258)	0.557(0.087)	0.080(0.018)
AR(0.2)	LARE	0.832(0.029)	18.63(3.286)	0.696(0.076)	0.093(0.019)
	LPRE	0.850(0.022)	18.15(4.075)	0.616(0.082)	0.083(0.012)
	LAD	0.835(0.028)	19.49(2.958)	0.583(0.127)	0.082(0.028)
	LS	0.858(0.021)	20.52(3.063)	0.587(0.111)	0.076(0.018)
AR(0.8)	LARE	0.913(0.031)	9.610(2.219)	0.833(0.128)	0.068(0.023)
	LPRE	0.889(0.025)	8.732(2.770)	0.857(0.105)	0.052(0.016)
	LAD	0.900(0.030)	15.85(2.980)	0.736(0.124)	0.072(0.026)
	LS	0.895(0.026)	14.60(2.970)	0.732(0.108)	0.007(0.029)
Band1	LARE	0.850(0.028)	14.23(3.010)	0.714(0.082)	0.097(0.020)
	LPRE	0.856(0.023)	15.64(3.274)	0.624(0.120)	0.083(0.016)
	LAD	0.844(0.030)	20.94(3.371)	0.543(0.114)	0.077(0.024)
	LS	0.856(0.023)	19.70(2.899)	0.626(0.090)	0.076(0.011)
Band2	LARE	0.868(0.032)	13.06(3.513)	0.782(0.163)	0.098(0.025)
	LPRE	0.864(0.030)	12.23(3.713)	0.763(0.148)	0.057(0.024)
	LAD	0.870(0.029)	17.23(3.555)	0.680(0.128)	0.073(0.027)
	LS	0.869(0.033)	16.46(2.470)	0.704(0.093)	0.071(0.010)

In each cell, mean (sd) based on 200 replicates

Table 2 Analysis of lung cancer data with LARE: main genetic effects and G×E interactions

		Interactions			
Gene	Main effects	Age	Gender	Smoking pack year	Smoking history
ADORA2B	−0.231				−0.260(0.76)
AKIRIN2	−0.281				
ASB12	−0.241				
C5ORF45	−0.042				
C14ORF93	−0.472				
C16ORF93	−0.160	−0.293(0.91)			
CAND1	0.309	−2.181(0.95)			
CBWD2	0.234				
CDR2	0.210				
CIAPIN1	0.187			−0.179(0.85)	
DCP1B	0.448				
DYRK2	−1.41	0.758(0.66)			
EIF4EBP1	0.081	−0.001(0.81)			
EMB	0.224				
FDXR	0.293			−0.477(0.99)	
GALK2	−0.158			−0.240(0.75)	
GOLGA7	−0.146	−0.096(0.45)			
HERPUD2	0.121				
HOXC13	−0.248	−0.145(0.98)			
ING1	−2.117				2.154(0.97)
INO80B	−0.164			−1.607(0.95)	
KIF21B	−0.391	−0.446(0.99)			
KLHDC1	−0.011			0.382(0.98)	
LIG4	−0.584		0.299(0.80)		
LINC00471	0.236				0.114(0.94)
LINC00476	0.258				0.056(0.55)
LRRC45	−0.136	−0.083(0.93)			
MCAT	0.103			0.180(0.96)	
MVD	−0.348				
NCALD	0.376		−0.605(0.70)		
OTUD1	0.189			0.038(0.34)	
PEX19	−0.444				0.045(0.55)
PHLPP1	−0.439				
PNPLA2	−0.193	0.014(0.55)			
PPM1A	−0.124			0.166(0.89)	
PPP2R2D	0.157	−0.234(0.67)			
RBM11	0.032		−0.291(0.71)		
RNF6	−0.215	0.199(0.90)			
RNF126P1	0.225				
RPS27	0.134				−0.155(0.22)
SCAND2P	−0.002			0.329(0.35)	

(continued)

Table 2 (continued)

		Interactions			
Gene	Main effects	Age	Gender	Smoking pack year	Smoking history
SERTAD4	−0.356		0.350(0.91)		
SGSM3	0.285			−0.039(0.46)	
SH3RF1	−0.096				
SLC25A2	−0.009		−0.335(0.94)		
SPCS3	−0.310				0.340(0.66)
SPRED2	−0.260				
SRRM3	−0.317				−0.244(0.70)
TXN2	−0.339			0.012(0.46)	
UBE4B	0.418	−0.497(0.53)			
VPS13B	0.065			−0.108(0.99)	
ZNF727	0.401	−0.254(0.78)			

For the interactions, values in "()" are the stability results

adopted in the literature. A total of 819 gene expressions are included in downstream analysis. For each gene expression, we normalize to have mean 0 and variance 1.

We apply the proposed approach and select the optimal λ using fivefold cross validation. The detailed identification and estimation results are presented in Tables 2 (LARE) and 5 (LPRE, Appendix). As previously described, it is possible that the main effects corresponding to the identified interactions are not identified. To respect the "main effects, interactions" hierarchy, we add back such main effects and re-fit. Beyond the proposed, we also apply the LS and LAD methods. The summary of applying different methods is provided in Table 4 (Appendix). Detailed estimation and identification results using the alternatives are presented in Tables 6 and 7 (Appendix). Different methods identify different sets of main effects and interactions. It is interesting that all of the main effects and interactions identified by LPRE are identified by LARE. They may represent more reliable findings. The LAD method identifies much fewer effects.

To complement the identification and estimation analysis, we evaluate stability. Specifically, we randomly remove ten subjects and then analyze data. This procedure is repeated 200 times. We then compute the probability that an interaction term is identified. Such an evaluation has been conducted in the literature. The stability results are provided in Tables 2, 5, 6, and 7 (Appendix). We can see that most of the identified interactions are relatively stable, with many having probabilities of being identified close to one.

Table 3 Summary of Simulation I. In each cell, mean (sd) based on 200 replicates

		AUC	SE	TPR	FPR
$\log(\varepsilon) \sim N(0, 1)$					
Independent	LARE	0.835(0.033)	10.58(1.742)	0.639(0.085)	0.092(0.018)
	LPRE	0.837(0.032)	11.64(1.918)	0.603(0.077)	0.080(0.011)
	LAD	0.848(0.033)	10.63(1.832)	0.599(0.089)	0.076(0.019)
	LS	0.836(0.031)	10.15(1.872)	0.590(0.089)	0.079(0.019)
AR(0.2)	LARE	0.859(0.038)	9.522(2.475)	0.697(0.135)	0.071(0.024)
	LPRE	0.858(0.036)	11.13(2.080)	0.660(0.126)	0.064(0.024)
	LAD	0.877(0.033)	9.363(2.230)	0.672(0.109)	0.079(0.015)
	LS	0.858(0.033)	8.972(1.929)	0.661(0.112)	0.066(0.018)
AR(0.8)	LARE	0.920(0.032)	5.834(2.577)	0.844(0.161)	0.057(0.028)
	LPRE	0.922(0.032)	6.932(2.116)	0.833(0.122)	0.039(0.028)
	LAD	0.939(0.025)	5.268(1.744)	0.835(0.127)	0.051(0.027)
	LS	0.923(0.028)	5.598(1.744)	0.795(0.127)	0.036(0.029)
Band1	LARE	0.860(0.033)	9.793(2.599)	0.721(0.136)	0.088(0.020)
	LPRE	0.860(0.033)	9.338(1.871)	0.698(0.118)	0.064(0.019)
	LAD	0.883(0.028)	8.455(1.623)	0.690(0.111)	0.072(0.016)
	LS	0.862(0.031)	8.573(2.849)	0.674(0.141)	0.060(0.023)
Band2	LARE	0.904(0.028)	6.907(2.262)	0.784(0.172)	0.072(0.021)
	LPRE	0.893(0.033)	6.706(2.842)	0.741(0.138)	0.052(0.021)
	LAD	0.915(0.033)	6.638(2.131)	0.757(0.142)	0.058(0.022)
	LS	0.899(0.034)	6.984(2.112)	0.746(0.147)	0.049(0.028)
$\log(\varepsilon) \sim \mathit{Unif}(-2, 2)$					
Independent	LARE	0.830(0.034)	12.13(2.716)	0.647(0.092)	0.082(0.016)
	LPRE	0.841(0.036)	10.64(2.592)	0.597(0.118)	0.070(0.021)
	LAD	0.849(0.028)	10.74(1.886)	0.570(0.136)	0.079(0.025)
	LS	0.835(0.029)	11.15(2.094)	0.540(0.135)	0.069(0.027)
AR(0.2)	LARE	0.846(0.027)	9.231(1.574)	0.657(0.139)	0.088(0.018)
	LPRE	0.854(0.031)	10.91(2.148)	0.628(0.120)	0.076(0.025)
	LAD	0.872(0.031)	9.846(1.416)	0.628(0.127)	0.072(0.024)
	LS	0.852(0.034)	9.721(1.772)	0.599(0.153)	0.062(0.026)
AR(0.8)	LARE	0.923(0.030)	6.454(1.449)	0.793(0.150)	0.048(0.026)
	LPRE	0.921(0.029)	7.832(1.893)	0.792(0.180)	0.035(0.032)
	LAD	0.934(0.027)	6.007(1.452)	0.798(0.123)	0.054(0.022)
	LS	0.917(0.027)	6.932(2.101)	0.772(0.207)	0.034(0.030)
Band1	LARE	0.851(0.033)	9.944(1.521)	0.683(0.117)	0.079(0.028)
	LPRE	0.847(0.033)	9.832(1.931)	0.658(0.108)	0.075(0.039)
	LAD	0.878(0.029)	9.126(1.865)	0.694(0.117)	0.080(0.020)
	LS	0.857(0.033)	9.465(1.816)	0.662(0.121)	0.091(0.032)
Band2	LARE	0.877(0.034)	6.803(1.303)	0.797(0.171)	0.070(0.023)
	LPRE	0.889(0.032)	6.943(1.503)	0.763(0.129)	0.058(0.023)
	LAD	0.911(0.032)	6.439(1.458)	0.760(0.148)	0.062(0.021)
	LS	0.893(0.027)	6.498(1.860)	0.733(0.162)	0.057(0.027)

Table 4 Analysis of lung cancer data using different methods: the numbers of identified main effects and interactions and overlaps

	LARE	LPRE	LAD	LS
LARE	52/38	43/32	8/3	23/12
LPRE		43/32	7/3	20/10
LAD			34/13	14/6
LS				45/32

In each cell, number of identified main effects/number of identified interactions

5 Discussion

The identification of important G×E interactions remains a challenging problem. In this article, we have introduced using the relative error criteria as loss functions. A penalized approach has been adopted for estimation and selection. Simulation shows that the proposed approach has performance comparable to or better than the alternatives. Thus it may be provide a useful alternative for data analysis. A limitation of this study is that the asymptotic properties have not been established. In the analysis of a lung cancer dataset, the LARE and LPRE results are relatively consistent but different from the alternatives. The identified interactions are reasonably stable. More examination of the findings is needed in the future.

Acknowledgements We thank the participants of Joint 24th ICSA Applied Statistics Symposium and 13th Graybill Conference in Colorado and organizers of this proceedings. This study was supported in part by the National Science Foundation of China (Grant No. 11401561), National Social Science Foundation of China (13CTJ001, 13&ZD148), NIH (CA016359, CA191383), and the U.S. VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development.

Appendix

See Tables 2, 4, 5, 6, and 7.

Table 5 Analysis of lung cancer data with LPRE: main genetic effects and G×E interactions

		Interactions			
Probe	Main effects	Age	Gender	Smoking pack year	Smoking history
ADORA2B	0.548				−1.093(0.73)
AKIRIN2	−0.287				
ASB12	−0.156				
C14ORF93	−0.347				
C16ORF93	−0.023	−0.290(0.85)			
CAND1	0.536	−2.113(0.98)			
CBWD2	0.293				
CDR2	0.200				
CIAPIN1	0.163			−0.369(0.88)	
DCP1B	0.482				
DYRK2	−1.473	0.523(0.57)			
FDXR	0.236			−0.655(0.98)	
GALK2	−0.021			−0.213(0.58)	
GOLGA7	0.037	−0.091(0.41)			
HERPUD2	0.065				
HOXC13	−0.064	−0.173(0.98)			
ING1	−1.133				0.939(0.96)
INO80B	−0.100			−0.803(0.75)	
KIF21B	−0.176	−0.330(0.98)			
KLHDC1	0.030			0.232(0.81)	
LIG4	−0.446		0.088(0.96)		
LINC00471	0.037				0.146(0.71)
LRRC45	−0.249	−0.234(0.86)			
MCAT	0.017			0.082(0.65)	
MVD	−0.321				
NCALD	0.057		−0.436(0.61)		
OTUD1	0.131			0.094(0.32)	
PEX19	−0.535				0.128(0.36)
PHLPP1	−0.355				
PNPLA2	−0.073	0.009(0.18)			
PPP2R2D	0.137	−0.173(0.62)			
RBM11	0.119		−0.263(0.46)		
RNF6	−0.256	0.303(0.85)			
SERTAD4	−0.040		0.148(0.60)		
SGSM3	0.176			−0.004(0.16)	
SH3RF1	−0.177				
SLC25A2	0.040		−0.314(0.78)		
SPCS3	−0.559				0.351(0.42)
SPRED2	−0.437				
SRRM3	0.337				−0.744(0.80)
TXN2	−0.225			0.160(0.32)	
UBE4B	0.327	−0.524(0.50)			
VPS13B	0.188			−0.406(0.91)	

For the interactions, values in "()" are the stability results

Table 6 Analysis of lung cancer data with LAD: main genetic effects and G×E interactions

		Interactions			
Probe	Main effects	Age	Gender	Smoking pack year	Smoking history
ADORA2B	−0.053				
AKR1A1	−0.085				
ALG9	−0.072				
ANKRD54	0.012				
ANP32B	−0.040				
ARFGAP2	−0.034				
ARL15	0.029				
ASB12	−0.014				
ASCC1	0.003				
ATP8B2	0.023				
C2ORF16	−0.032		−0.032(0.37)		0.029(0.59)
C2ORF42	0.055				
VIPAS39	0.044	−0.030(0.59)			
C16ORF93	−0.011	−0.084(0.86)			
CAND1	−0.001	−0.362(1.00)			
CD46	−0.001				
CHKA	−0.041				
DCP1B	0.050				0.036(0.87)
DNAJC21	0.035				
DPY19L1	0.030	−0.026(0.69)			
DUSP6	−0.007				
EIF3F	−0.009				
EMB	−0.157				
FCRLB	−0.050				
FDXR	−0.009			−0.159(0.96)	
GABPA	−0.095				
GINS4	−0.069				
HKR1	−0.011				−0.008(0.13)
KLF10	−0.028	0.087(0.73)			
LIN37	0.038	−0.013(0.97)			
LINC00515	0.029	−0.043(0.97)			
PAF1	−0.053				
SPRED2	−0.164				0.046(0.77)
TWISTNB	0.079				

For the interactions, values in “()” are the stability results

Table 7 Analysis of lung cancer data with LS: main genetic effects and G×E interactions

		Interactions			
Probe	Main effects	Age	Gender	Smoking pack year	Smoking history
ADORA2B	−0.027				
ARL15	0.012		−0.046(0.69)		
C2ORF16	0.005		−0.013(0.53)		
C11ORF52	0.029		−0.032(0.49)		
C14ORF93	−0.112				
C16ORF93	−0.017	−0.096(0.99)			
CAND1	−0.035	−0.347(0.97)			
CBWD2	0.000		0.027(0.58)		
CCDC171	0.011			−0.026(0.92)	
CDR2	−0.003				
DCP1B	0.103				
DNAJB13	0.045				
DNAJC30	−0.024			0.025(0.85)	
DYRK1B	0.011	−0.030(0.58)			
EIF3F	−0.008				
EIF4EBP1	−0.007	−0.025(0.97)			
EMB	−0.081				
FDXR	−0.009			−0.135(1.00)	
GEMIN8	0.036		−0.045(0.57)		
HIST1H2AJ	−0.002	−0.013(0.25)			
HNRNPDL	−0.019				−0.010(0.51)
HOXC13	−0.006	−0.003(0.43)			
ING1	0.016		−0.031(0.36)		
INO80B	−0.004			−0.282(0.96)	
KLHDC1	0.000			0.042(0.79)	
KLHL7	0.015				
LIN37	0.014	−0.050(0.63)			
LINC00515	0.028	−0.037(0.92)			
LRRC45	0.016	−0.006(0.41)		0.024(0.81)	
MVD	−0.050				
PAF1	0.054		−0.125(0.84)		
PHLPP1	−0.034				
PIK3CB	−0.032				−0.007(0.76)
PNPLA2	−0.014				
POLN	−0.011				0.030(0.64)
PPHLN1	0.047				
RNF6	0.014	0.007(0.92)			
RPS27	0.027				−0.092(0.76)

(continued)

Table 7 (continued)

		Interactions			
Probe	Main effects	Age	Gender	Smoking pack year	Smoking history
SGSM3	−0.003			−0.020(0.37)	
SPRED2	−0.047				
SYNCRIP	−0.042	0.029(0.28)			
TRAM1L1	0.002		−0.044(0.81)		
TWISTNB	0.031				
UBE4B	0.021	−0.098(0.64)			
ZNF737	−0.038			0.007(0.19)	

For the interactions, values in "()" are the stability results

References

Bien, J., Taylor, J., Tibshirani, R., et al.: A lasso for hierarchical interactions. The Annals of Statistics **41**(3), 1111–1141 (2013)

Caspi, A., Moffitt, T.E.: Gene–environment interactions in psychiatry: joining forces with neuroscience. Nature Reviews Neuroscience **7**(7), 583–590 (2006)

Chen, K., Guo, S., Lin, Y., Ying, Z.: Least absolute relative error estimation. Journal of the American Statistical Association **105**(491), 1104–1112 (2010)

Chen, K., Lin, Y., Wang, Z., Ying, Z.: Least product relative error estimation. arXiv preprint arXiv:1309.0220 (2013)

Cordell, H.J.: Detecting gene–gene interactions that underlie human diseases. Nature Reviews Genetics **10**(6), 392–404 (2009)

Hunter, D.J.: Gene–environment interactions in human diseases. Nature Reviews Genetics **6**(4), 287–298 (2005)

Hunter, D.R., Li, R.: Variable selection using mm algorithms. Annals of statistics **33**(4), 1617–1642 (2005)

Khoshgoftaar, T.M., Bhattacharyya, B.B., Richardson, G.D.: Predicting software errors, during development, using nonlinear regression models: a comparative study. Reliability, IEEE Transactions on **41**(3), 390–395 (1992)

Li, Z., Lin, Y., Zhou, G., Zhou, W.: Empirical likelihood for least absolute relative error regression. Test **23**(1), 86–99 (2014)

Liu, J., Huang, J., Zhang, Y., Lan, Q., Rothman, N., Zheng, T., Ma, S.: Identification of gene–environment interactions in cancer studies using penalization. Genomics **102**(4), 189–194 (2013)

North, K.E., Martin, L.J.: The importance of gene-environment interaction implications for social scientists. Sociological Methods & Research **37**(2), 164–200 (2008)

Park, H., Stefanski, L.: Relative-error prediction. Statistics & probability letters **40**(3), 227–236 (1998)

Shi, X., Liu, J., Huang, J., Zhou, Y., Xie, Y., Ma, S.: A penalized robust method for identifying gene–environment interactions. Genetic epidemiology **38**(3), 220–230 (2014)

Thomas, D.: Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annual review of public health **31**, 21–36 (2010)

Tsionas, E.G.: Bayesian analysis of least absolute relative error regression. Communications in Statistics-Theory and Methods **43**(23), 4988–4997 (2014)

Van Dam, L.C., Ernst, M.O.: Relative errors can cue absolute visuomotor mappings. Experimental brain research **233**(12), 3367–3377 (2015)

Wu, C., Cui, Y., Ma, S.: Integrative analysis of gene–environment interactions under a multi-response partially linear varying coefficient model. Statistics in medicine **33**(28), 4988–4998 (2014)

Wu, C., Ma, S.: A selective review of robust variable selection with applications in bioinformatics. Briefings in bioinformatics **16**, 873–883 (2015)

Wu, T.T., Lange, K.: Coordinate descent algorithms for lasso penalized regression. The Annals of Applied Statistics pp. 224–244 (2008)

Xia, X., Liu, Z., Yang, H.: Regularized estimation for the least absolute relative error models with a diverging number of covariates. Computational Statistics & Data Analysis (2015)

Zhang, C.: Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics pp. 894–942 (2010)

Zhang, Q., Wang, Q.: Local least absolute relative error estimating approach for partially linear multiplicative model. Statistica Sinica **23**, 1091–1116 (2013)

Zhu, R., Zhao, H., Ma, S.: Identifying gene–environment and gene–gene interactions using a progressive penalization approach. Genetic epidemiology **38**(4), 353–368 (2014)

Zimmermann, P., Brückl, T., Nocon, A., Pfister, H., Binder, E.B., Uhr, M., Lieb, R., Moffitt, T.E., Caspi, A., Holsboer, F., et al.: Interaction of fkbp5 gene variants and adverse life events in predicting depression onset: results from a 10-year prospective community study. American Journal of Psychiatry **168**, 1107–1116 (2011)

Partially Supervised Sparse Factor Regression For Multi-Class Classification

Chongliang Luo, Dipak Dey, and Kun Chen

Abstract The classical linear discriminant analysis (LDA) may perform poorly in multi-class classification with high-dimensional data. We propose a partially supervised sparse factor regression (PSFAR) approach, to jointly explore the potential low-dimensional structures in the high-dimensional class mean vectors and the common covariance matrix required in LDA. The problem is formulated as a multivariate regression analysis, with predictors constructed from the class labels and responses from the high-dimensional features. The regression coefficient matrix is then composed of the class means, for which we explore a sparse and low rank structure; we further explore a parsimonious factor analysis representation in the covariance matrix. As such, our model assumes that the high-dimensional features are best separated in their means in a low-dimensional subspace, subject to a few unobserved latent factors. We propose a regularized log-likelihood criterion for model estimation, for which an efficient Expectation-Maximization algorithm is developed. The efficacy of PSFAR is demonstrated by both simulation studies and a real application using handwritten digit data.

Keywords Factor analysis • Linear discriminant analysis • Multi-class classification • Reduced-rank regression • Variable selection

1 Introduction

Given a collection of features from different classes/categories, classification analysis aims to understand the discrepancy between the classes and to construct a classifier to predict the class memberships of future observations. Several commonly-used classification methods include linear discriminant analysis (LDA) (Fisher 1936), logistic regression, and support vector machine (SVM) (Vapnik and

C. Luo • D. Dey • K. Chen (✉)
Department of Statistics, University of Connecticut, Storrs, CT, USA

Center for Public Health and Health Policy, University of Connecticut Health Center, Farmington, CT, USA
e-mail: kun.chen@uconn.edu

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_24

Vapnik 1998). Coming to the big data era, classification tasks with high-dimensional features are frequently encountered, and often the number of classes is also large, resulting in high-dimensional multi-class classification problems. For example, in some image classification tasks, there may be a large number of image types, and the number of features, e.g., the number of pixels in each image, can also be high (LeCun et al. 1998). Based upon existing binary classifiers, several multi-class classification techniques such as one-versus-the-rest or one-versus-another were studied (Allwein et al. 2001). Friedman et al. (2010) studied multinomial logistic regression and its regularized versions to handle high dimensionality. See Li et al. (2006), Lorena et al. (2009) and Zhang and Zhou (2014) for some comprehensive reviews of the existing multi-class classification methods.

Here we mainly focus on using LDA for multi-class classification, due to its simplicity and wide applicability. The method relies on the assumption that the data from different classes follow Gaussian distributions with different means and a common covariance matrix. With high-dimensional data, feature selection and regularized estimation are desired and often required, because otherwise the classification based on sample estimates can be as bad as random guessing (Bickel and Levina 2004; Fan and Fan 2008). Fan and Fan (2008) thoroughly discussed the impact of high dimensionality in classification and provided a remedy termed "features annealed independence rule", combining the ideas of feature screening and the independence rule proposed by Bickel and Levina (2004). Several other sparse LDA methods were developed, e.g., Shao et al. (2011) and Cai and Liu (2011). LDA is closely related to multivariate regression and canonical correlation analysis (CCA) (Friedman et al. 2001; Reinsel and Velu 1998), in which the task is formulated as examining the associations between the dummy variables from the class labels and the high-dimensional features. Motivated by these connections, we argue that the high-dimension LDA can be alternatively achieved through a regularized multivariate regression analysis. The dimension of the multivariate problem is determined by both the number of features and the number of classes. As such, the reformulation is particularly useful when there are both a large number of classes and a large number of features, as various multivariate dimension reduction techniques then become effective. The regularized multivariate analysis has undergone exciting development in recent years. Reduced rank regression (RRR) achieved dimension reduction through restricting the rank of the coefficient matrix (Anderson 1951; Reinsel and Velu 1998; Izenman 2008; Bunea et al. 2011; Chen et al. 2013), and recently many works considered the incorporation of the reduced-rank representation with other low-dimensional structures such as sparsity (Chen et al. 2012; Chen and Huang 2012; Bunea et al. 2012). Joint mean and covariance estimation has also been studied (Rothman et al. 2010; Chen and Huang 2016).

We propose a partially supervised sparse factor regression (PSFAR) approach, to jointly explore and estimate the potential low-dimensional structures in the high-dimensional class means and the common covariance matrix. As these quantities are essential in LDA, a better estimation of them that is suitable in high-dimensional scenarios has great potential in improving the classification performance. In Sect. 2, the problem is formulated as a multivariate regression analysis, with predictors constructed from the class labels and responses from the high-dimensional features. The regression coefficient matrix is then composed of the class means, for which we explore a sparse and low rank structure, to jointly achieve feature selection and subspace recovery. We further explore a parsimonious factor analysis representation of the covariance matrix. As such, our model assumes that the high-dimensional features are best separated in their means in a sparse and low-dimensional subspace, subject to a few unobserved common latent factors. In Sect. 3, we propose a regularized log-likelihood criterion for model estimation, for which an efficient Expectation-Maximization (EM) algorithm is developed. In Sect. 4, we conduct simulation studies to demonstrate that PSFAR can greatly improve the classical LDA. We apply the proposed method to analyze MNIST handwritten digits data in Sect. 5.

2 A Sparse Factor Regression Model for Multi-Class Classification

2.1 LDA and Its Connections with Multivariate Regression

Suppose we observe $\mathbf{y}_i \in \mathbb{R}^q, i = 1, 2, \ldots, n$, consisting of n_k observations from each of the K classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$, with $n = \sum_{k=1}^K n_k$. We assume that the observations from each class follow Gaussian distribution with a common covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathbf{y}_i \sim \mathbb{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, for $\mathbf{y}_i \in \mathcal{C}_k, k = 1, \ldots, K$ (Friedman et al. 2001). Without loss of generality, we assume all the samples are centered, i.e. $\overline{\mathbf{y}} = \sum_{i=1}^n \mathbf{y}_i/n = 0$. Denote the sample mean vector of class $\mathcal{C}_k$ as $\overline{\mathbf{y}}_k = \sum_{\mathbf{y}_i \in \mathcal{C}_k} \mathbf{y}_i/n_k$.

LDA looks for "discriminant directions" $\phi_j \in \mathbb{R}^q, j = 1, \ldots, K-1$, along which the projected observations tend to cluster together when they are from the same class and tend to be separated from each other otherwise (Fisher 1936; Rao 1948). Given the data, the sample LDA is conducted by maximizing the "Rayleigh coefficient",

$$\frac{|\boldsymbol{\Phi}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_b\boldsymbol{\Phi}|}{|\boldsymbol{\Phi}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_w\boldsymbol{\Phi}|},$$

where $\boldsymbol{\Phi} = [\phi_1, \ldots, \phi_{K-1}] \in \mathbb{R}^{q\times(K-1)}$,

$$\widehat{\boldsymbol{\Sigma}}_w = \frac{1}{n-K}\sum_{k=1}^K \sum_{y_i \in \mathcal{C}_k} (\mathbf{y}_i - \overline{\mathbf{y}}_k)(\mathbf{y}_i - \overline{\mathbf{y}}_k)^{\mathrm{T}}$$

is the within-class covariance matrix, and

$$\widehat{\boldsymbol{\Sigma}}_b = \frac{1}{K-1}\sum_{k=1}^K n_k(\overline{\mathbf{y}}_k - \overline{\mathbf{y}})(\overline{\mathbf{y}}_k - \overline{\mathbf{y}})^{\mathrm{T}}$$

is the between-class covariance matrix. That is, the discriminant directions are found by maximizing the ratio between the determinant of the within-class covariance matrix of the projected observations and that of the between-class covariance matrix of the projected observations. This is a generalized eigenvalue problem, and a set of solution is given by the $K-1$ eigenvectors of $\widehat{\mathbf{\Sigma}}_w^{-1}\widehat{\mathbf{\Sigma}}_b$. When $\widehat{\mathbf{\Sigma}}_w$ is singular, a common practice is to replace it by $\widehat{\mathbf{\Sigma}}_w + \delta\mathbf{I}$, where δ is a small positive constant (McLachlan 2004). To classify an observation $\mathbf{y} \in \mathbb{R}^q$, the distances from its projection $\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}$ to the projected class centers are computed, and it is classified to $\mathcal{C}_{\widehat{k}}$ with $\widehat{k} = \arg\min_k \|\mathbf{\Phi}^{\mathrm{T}}\mathbf{y} - \mathbf{\Phi}^{\mathrm{T}}\overline{\mathbf{y}}_k\|$.

The classification problem here can also be understood through a regression setup. For any $\mathbf{y}_i \in \mathcal{C}_k$, let $\mathbf{x}_i \in \mathbb{R}^K$ be a vector of 0's except the k^{th} element being 1. Then the model can be written as

$$\mathbf{y}_i = \mathbf{C}^{\mathrm{T}}\mathbf{x}_i + \mathbf{e}_i, \qquad i = 1, \ldots, n. \tag{1}$$

where $\mathbf{C} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K]^{\mathrm{T}} \in \mathbb{R}^{K\times q}$, and $\mathbf{e}_i \in \mathbb{R}^q$ are the random error vectors following $\mathbb{N}(\mathbf{0}, \mathbf{\Sigma})$. Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times q}$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times K}$, and $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times q}$. Then in matrix form, the model becomes

$$\mathbf{Y} = \mathbf{X}\mathbf{C} + \mathbf{E}.$$

The design matrix satisfies $\mathbf{D} = \mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{diag}(n_1, \ldots, n_K)$. Glahn (1968) showed that when $n > q + K$, finding the discriminant directions in LDA is exactly the same as finding the canonical directions in the CCA analysis of $\mathbf{Y}$ and $\mathbf{X}$. Interestingly, the CCA itself, can be formulated as a special case of reduced-rank regression (RRR) (Izenman 1975; Reinsel and Velu 1998), for which the estimation criterion is

$$\min_{\mathbf{C}} \mathrm{tr}\{(\mathbf{Y} - \mathbf{X}\mathbf{C})\mathbf{\Gamma}(\mathbf{Y} - \mathbf{X}\mathbf{C})^{\mathrm{T}}\}, \tag{2}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of the enclosed matrix, $\mathbf{C} = \mathbf{A}\mathbf{B}^{\mathrm{T}}$, with $\mathbf{A} \in \mathbb{R}^{K\times r}$, $\mathbf{B} \in \mathbb{R}^{q\times r}$ satisfying the constraints $(\mathbf{X}\mathbf{A})^{\mathrm{T}}(\mathbf{X}\mathbf{A}) = \mathbf{I}$ and $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I}$, and $\mathbf{\Gamma}$ is a given positive definite weighting matrix. Izenman (1975) showed that when $\mathbf{\Gamma} = (\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}$, the solution of $\mathbf{B}$ in (2) corresponds to the canonical directions in CCA or the discriminant directions $\mathbf{\Phi}$ in LDA. Here r is the model rank, which corresponds to the number of discriminant directions in LDA.

2.2 A Partially Supervised Sparse Factor Regression Model

When there are a large number of classes, it is often plausible to assume that the class centers live in a lower-dimensional space, which can be achieved by restricting the rank of the coefficient matrix $\mathbf{C}$ in (1), i.e., $r(\mathbf{C}) = r \leq \min(K, q)$. When the dimension of features q is large, there may exist many noisy features that do not contribute to the discrimination of the different classes, which can be achieved by assuming certain sparsity pattern in $\mathbf{C}$. For the covariance matrix $\mathbf{\Sigma}$, we follow the factor analysis model to assume that

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^{\mathrm{T}} + \mathbf{\Psi}, \tag{3}$$

where $\mathbf{L} \in \mathbb{R}^{q\times h}$ and $\mathbf{\Psi} = \mathbf{diag}(\psi_1,\ldots,\psi_q) \in \mathbb{R}^{q\times q}$ is a diagonal matrix with positive diagonal elements. This "low-rank + diagonal" structure provides a parsimonious way of parameterizing $\mathbf{\Sigma}$. The effective number of parameters used to determine $\mathbf{\Sigma}$ can be dramatically reduced when h is much smaller than q. This formulation seeks a few latent factors to explain the dependence structure of the random variables. As such, our method can be viewed as based on the following model as an extension of (1),

$$\mathbf{y}_i = \mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{x}_i + \mathbf{L}\mathbf{z}_i + \tilde{\mathbf{e}}_i, \qquad i = 1,\ldots,n, \tag{4}$$

where $\mathbf{z}_i \sim \mathbb{N}_h(\mathbf{0},\mathbf{I})$ are independent latent factors, $\mathbf{L} \in \mathbb{R}^{q\times h}$ is the factor loading matrix, $\tilde{\mathbf{e}}_i \sim \mathbb{N}_q(\mathbf{0},\mathbf{\Psi})$ are independent random error vectors, and $\mathrm{cov}(\mathbf{z}_i,\tilde{\mathbf{e}}_i) = \mathbf{0}$. Write $\mathbf{A} = [\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_K]^{\mathrm{T}}, \boldsymbol{\alpha}_k \in \mathbb{R}^r$, then $\mathbf{B}\boldsymbol{\alpha}_k$ gives the mean of $\mathcal{C}_k$. Feature selection can then be achieved by exploring the sparsity pattern of $\mathbf{B}$. We remark that certain constraints on $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{L}$ are needed for them to be identifiable (up to some orthogonal rotations), which will be discussed later.

In the regression model (4), $\mathbf{A}^{\mathrm{T}}\mathbf{x}_i$ can be regarded as some supervised factors, as they are constructed as some linear combinations of the predictors $\mathbf{x}_i$ (class labels) and are fully observable. In contrast, $\mathbf{z}_i$'s are some unsupervised latent factors and are completely unobservable. The dependency structures among the responses (observed features) can only be fully understood by exploring both types of factors jointly. Conditional on the supervised and unsupervised factors, the elements of $\mathbf{y}_i$ then become independent to each other. We thus name the proposed model (4) as a *partially supervised sparse factor regression* model (PSFAR). We utilize PSFAR to perform high-dimensional multi-class LDA. That is, we first fit PSFAR to get $\widehat{\mathbf{A}} = [\widehat{\boldsymbol{\alpha}}_1,\ldots,\widehat{\boldsymbol{\alpha}}_K]^{\mathrm{T}}$, $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{\Sigma}}$; these estimated class centers and covariance matrix are then used in LDA to perform classification.

3 Regularized Estimation

3.1 Penalized Log-Likelihood Criterion

The log-likelihood function from model (4), up to some constant, is

$$\ell(\mathbf{\Theta};\mathbf{Y},\mathbf{X}) = -\frac{1}{n}\left[\log|\mathbf{L}\mathbf{L}^{\mathrm{T}} + \mathbf{\Psi}| + \mathrm{tr}\{(\mathbf{L}\mathbf{L}^{\mathrm{T}} + \mathbf{\Psi})^{-1}\mathbf{S}\}\right],$$

where $\mathbf{\Theta} = \{\mathbf{A}, \mathbf{B}, \mathbf{L}, \mathbf{\Psi}\}$, and

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{x}_i)(\mathbf{y}_i - \mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}.$$

We propose to maximize the following penalized log-likelihood criterion for conducting model estimation,

$$\ell(\mathbf{\Theta}; \mathbf{Y}, \mathbf{X}) - \lambda\rho(\mathbf{B}), \tag{5}$$

subject to $\mathbf{A}\mathbf{D}\mathbf{A} = \mathbf{I}$, where recall that $\mathbf{D} = \mathbf{diag}(n_1, \ldots, n_K)$. The orthogonal constraint is to ensure the identifiability of the model parameters. Here, $\rho(\cdot)$ is a sparsity-inducing penalty function, with λ being the tuning parameter. To conduct feature selection, we mainly consider the row-wise group lasso penalty (Yuan and Lin 2006), i.e.,

$$\rho(\mathbf{B}) = \|\mathbf{B}\|_{2,1} = \sum_{j=1}^{q}\sqrt{\sum_{l=1}^{r} b_{jl}^2}.$$

3.2 EM Algorithm

We develop an EM algorithm to efficiently optimize the criterion in (5), by treating the latent factors as missing data. Denote the unobserved latent factors as $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]^{\mathrm{T}}$. Since $\mathbf{y}_i \mid \mathbf{x}_i \sim \mathbb{N}(\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{x}_i + \mathbf{L}\mathbf{z}_i, \mathbf{\Psi})$ and $\mathbf{z}_i \sim \mathbb{N}(0, \mathbf{I})$, the complete penalized log-likelihood given $(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$, up to some constant, is

$$\begin{aligned}\ell_c(\mathbf{\Theta}; \mathbf{Y}, \mathbf{X}, \mathbf{Z}) = &-\frac{1}{n^2}\sum_{i=1}^{n}\{\mathbf{z}_i^{\mathrm{T}}\mathbf{z}_i + (\mathbf{y}_i - \mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{x}_i - \mathbf{L}\mathbf{z}_i)^{\mathrm{T}}\mathbf{\Psi}^{-1}(\mathbf{y}_i - \mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{x}_i - \mathbf{L}\mathbf{z}_i)\} \\ &-\frac{1}{n}\log|\mathbf{\Psi}| - \lambda\rho(\mathbf{B}).\end{aligned}$$

In the E-step of the EM algorithm, we compute the conditional expectation of the complete log-likelihood given the observed data and the current parameter estimates $\mathbf{\Theta}_c$, i.e., $Q(\mathbf{\Theta}; \mathbf{\Theta}_c) = \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y}}\ell_c(\mathbf{\Theta}; \mathbf{Y}, \mathbf{X}, \mathbf{Z})$. By using the fact that $\mathbf{z}_i \mid \mathbf{y}_i \sim \mathbb{N}(\mathbf{L}_c^{\mathrm{T}}\mathbf{\Sigma}_c^{-1}(\mathbf{y}_i - \mathbf{B}_c\mathbf{A}_c^{\mathrm{T}}\mathbf{x}_i), (\mathbf{I} + \mathbf{L}_c^{\mathrm{T}}\mathbf{\Psi}_c^{-1}\mathbf{L}_c)^{-1})$, we have

$$\begin{aligned}Q(\mathbf{\Theta}; \mathbf{\Theta}_c) = &-\frac{1}{n}\mathrm{tr}\left(\mathbf{\Psi}^{-1}[\mathbf{L}\{(\mathbf{I} + \mathbf{L}_c^{\mathrm{T}}\mathbf{\Sigma}_c^{-1}\mathbf{L}_c)^{-1} + \mathbf{L}_c^{\mathrm{T}}\mathbf{\Sigma}_c^{-1}\mathbf{S}_c^{(1)}\mathbf{\Sigma}_c^{-1}\mathbf{L}_c\}\mathbf{L}^{\mathrm{T}}\right. \\ &\left.-2\mathbf{L}\mathbf{L}_c^{\mathrm{T}}\mathbf{\Sigma}_c^{-1}\mathbf{S}_c^{(2)} + \mathbf{S}]\right) \\ &-\frac{1}{n}\log|\mathbf{\Psi}| - \lambda\rho(\mathbf{B}),\end{aligned} \tag{6}$$

where $\boldsymbol{\Sigma}_c = \mathbf{L}_c\mathbf{L}_c^{\mathrm{T}} + \boldsymbol{\Psi}_c$, and

$$\mathbf{S}_c^{(1)} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{B}_c\mathbf{A}_c^{\mathrm{T}}\mathbf{x}_i)(\mathbf{y}_i - \mathbf{B}_c\mathbf{A}_c^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}},$$

$$\mathbf{S}_c^{(2)} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{B}_c\mathbf{A}_c^{\mathrm{T}}\mathbf{x}_i)(\mathbf{y}_i - \mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}.$$

In the M-step, we develop a block-wise coordinate descent algorithm to maximize $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}_c)$. For fixed $(\mathbf{A}, \mathbf{B})$, $\mathbf{L}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$ are updated as

$$\begin{aligned}
\widehat{\mathbf{L}} &= \mathbf{S}_c^{(2)}\boldsymbol{\Sigma}_c^{-1}\mathbf{L}_c\left\{(\mathbf{I} + \mathbf{L}_c^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\mathbf{L}_c)^{-1} + \mathbf{L}_c^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}\mathbf{S}_c^{(1)}\boldsymbol{\Sigma}_c^{-1}\mathbf{L}_c\right\}^{-1},\\
\widehat{\boldsymbol{\Psi}} &= \mathbf{diag}\left(\mathbf{S} - \mathbf{S}_c^{(2)}\boldsymbol{\Sigma}_c^{-1}\widehat{\mathbf{L}}\widehat{\mathbf{L}}^{\mathrm{T}}\right),\\
\widehat{\boldsymbol{\Sigma}} &= \widehat{\mathbf{L}}\widehat{\mathbf{L}}^{\mathrm{T}} + \widehat{\boldsymbol{\Psi}}.
\end{aligned} \tag{7}$$

On the other hand, for fixed $\mathbf{L} = \widehat{\mathbf{L}}$ and $\boldsymbol{\Psi} = \widehat{\boldsymbol{\Psi}}$, the problem becomes

$$\min_{\mathbf{A},\mathbf{B}} \frac{1}{n}\mathrm{tr}\left\{(\mathbf{Y}_c - \mathbf{X}\mathbf{A}\mathbf{B}^{\mathrm{T}})\widehat{\boldsymbol{\Psi}}^{-1}(\mathbf{Y}_c - \mathbf{X}\mathbf{A}\mathbf{B}^{\mathrm{T}})^{\mathrm{T}}\right\} + \lambda\rho(\mathbf{B}), \text{ subject to } \mathbf{A}^{\mathrm{T}}\mathbf{D}\mathbf{A} = \mathbf{I},$$

where

$$\mathbf{Y}_c = \mathbf{Y} - (\mathbf{Y} - \mathbf{X}\mathbf{A}_c\mathbf{B}_c^{\mathrm{T}})\widehat{\boldsymbol{\Psi}}^{-1}\widehat{\mathbf{L}}\widehat{\mathbf{L}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\Psi}}/2,$$

and $\rho(\mathbf{B}) = \|\mathbf{B}\|_{2,1}$. For fixed $\mathbf{B}$,

$$\widehat{\mathbf{A}} = \arg\min_{\mathbf{A}} \|\mathbf{Y}_c\widehat{\boldsymbol{\Psi}}^{-\frac{1}{2}} - \mathbf{X}\mathbf{A}\mathbf{B}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}^{-\frac{1}{2}}\|_F^2, \text{ subject to } \mathbf{A}^{\mathrm{T}}\mathbf{D}\mathbf{A} = \mathbf{I}. \tag{8}$$

This is a weighted Procrustes' problem (Gower and Dijksterhuis 2004). The solution is $\widehat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{V}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U}$ and $\mathbf{V}$ are respectively the left and right singular matrices of the matrix $\mathbf{B}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}^{-1}\mathbf{Y}_c^{\mathrm{T}}\mathbf{X}\mathbf{D}^{-\frac{1}{2}}$. For fixed $\mathbf{A} = \widehat{\mathbf{A}}$,

$$\begin{aligned}
\widehat{\mathbf{B}} &= \arg\min_{\mathbf{B}} \frac{1}{n}\|\mathbf{Y}_c\widehat{\boldsymbol{\Psi}}^{-\frac{1}{2}} - \mathbf{X}\widehat{\mathbf{A}}\mathbf{B}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}^{-\frac{1}{2}}\|_F^2 + \lambda\|\mathbf{B}\|_{2,1}\\
&= \arg\min_{\mathbf{B}} \sum_{j=1}^{q}\{\frac{1}{n}\|(\mathbf{Y}_c^{\mathrm{T}})_j - \mathbf{X}\widehat{\mathbf{A}}(\mathbf{B})_j\|^2 + \lambda\widehat{\psi}_j\|(\mathbf{B})_j\|\}.
\end{aligned} \tag{9}$$

Here $(\mathbf{M})_j$ represents the jth row vector of the enclosed matrix $\mathbf{M}$, and $\widehat{\psi}_j$ is the jth diagonal element of $\widehat{\boldsymbol{\Psi}}$. This optimization problem is separable in each $(\mathbf{B})_j$ and admits explicit solution,

$$(\widehat{\mathbf{B}})_j = \mathcal{S}((\mathbf{Y}_c^{\mathrm{T}}\mathbf{X}\widehat{\mathbf{A}})_j, n\lambda\widehat{\psi}_j/2), \qquad j = 1, \ldots, q,$$

where

$$\mathcal{S}(\mathbf{u}, \lambda) = \begin{cases} \frac{\|\mathbf{u}\|-\lambda}{\|\mathbf{u}\|}\mathbf{u} & \|\mathbf{u}\| \geq \lambda, \\ \mathbf{0} & \|\mathbf{u}\| < \lambda, \end{cases}$$

is a thresholding function for a vector $\mathbf{u}$.

It can be seen that all the subproblems in the M-step are easy to solve, and the objective function of the M-step in (6) is monotone non-decreasing along the iterations. The M-step can be either solved fully or partially, i.e., iteratively update $(\mathbf{L}, \boldsymbol{\Psi}, \mathbf{A}, \mathbf{B})$ according to (7)–(9) either until convergence or for only a few times. The latter approach leads to a generalized EM algorithm, which may converge even faster than the standard EM based on our limited experience. In practice, the initial values of $\mathbf{A}, \mathbf{B}$ are obtained from a reduced-rank regression analysis; the initial values of $\mathbf{L}$ and $\boldsymbol{\Psi}$ are then obtained from a factor analysis of the sample covariance matrix of the residuals.

The proposed algorithm works for fixed model ranks and tuning parameter. In practice, we need to choose the rank r, the number of latent factors h and the penalty parameter λ. Both r and h are usually small in real applications. For any fixed r and h, we choose 50 λ values equally spaced on the log scale, to produce a whole spectrum of possible sparsity patterns in $\mathbf{B}$. To determine the best model, we use fivefold cross-validation based on the classification performance of different models.

4 Simulation

In our simulation study, we consider $r = 2$, $h = 2$, $q \in \{200, 400\}$, $K \in \{5, 10\}$ and both balanced and unbalanced scenarios. For balanced data, the size of each class is 40; for unbalanced data, the sizes of the classes are (20, 20, 40, 60, 60) and (20, 20, 20, 20, 40, 40, 60, 60, 60, 60) for $K = 5$ and $K = 10$, respectively. The matrix $\mathbf{B}$ is set as a row-wise sparse matrix with the first $q_0 = 20$ rows being nonzero, and the nonzero entries are randomly generated from a uniform distribution on the set $[-1, -0.5] \cup [0.5, 1]$. To generate $\mathbf{A}$, we first generate a random matrix of the same size with its entries being random samples from a normal distribution, and is then transformed so that $\mathbf{A}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{A} = \mathbf{I}$. To generate $\mathbf{L}$, we first generate a matrix of the same size with its entries being random samples from the standard normal distribution, and then the matrix is orthogonalized to obtain $\mathbf{L}$ to make $\mathbf{L}^{\mathrm{T}}\mathbf{L} = \sigma_1^2\mathbf{I}_h$. We set $\boldsymbol{\Psi} = \sigma_2^2\mathbf{I}$. We test various pairs of σ_1^2 and σ_2^2, which indicate the strengths of the latent variables and the random errors, respectively. With the generated $\mathbf{L}$ and $\boldsymbol{\Psi}$, the $\mathbf{e}_i$ vectors are generated from $\mathbb{N}(\mathbf{0}, \mathbf{L}\mathbf{L}^{\mathrm{T}} + \boldsymbol{\Psi})$. Finally, the data matrix $\mathbf{Y}$ is generated based on model (1), i.e., $\mathbf{Y} = \mathbf{X}\mathbf{A}\mathbf{B}^{\mathrm{T}} + \mathbf{E}$. In each experiment, a test data with a size of three times that of the training data is also generated. The experiment is replicated 50 times under each setting.

We consider the classical LDA, the sparse multinomial regression (SMR), and our proposed PSFAR method. SMR has been implemented in the R package *glmnet*. As a benchmark, we also include an oracle procedure (ORE), which performs LDA using the true class means and true covariance matrix. We use the out-of-sample misclassification rate (MCR) evaluated based on the test data to measure the classification accuracy of the methods. For PSFAR and SMR, we also evaluate their variable selection performance: the true positive rate (TPR) is computed as the ratio of correctly selected variables among the true related variables, and the false positive rate (FPR) is computed as the ratio of falsely selected variables among the irrelevant variables.

In our simulation example, as we set $r = 2$, it is assumed that the true class centers live in a two-dimensional space. To visualize the simulated data, Fig. 1 shows two typical scatter plots of the data points projected to the first two discriminant directions based on the true model. Table 1 report the simulation results for the unbalanced case. (The balanced case is omitted as it delivers similar message.) As expected, LDA performs the worst, especially when the number of variables is large. SMR tends to miss useful variables, resulting in much lower TPR comparing to PSFAR. PSFAR performs well in variable selection, and its classification performance is close to that of the oracle procedure. In PSFAR, the r and h are almost always correctly selected by cross validation; we thus omit the results.

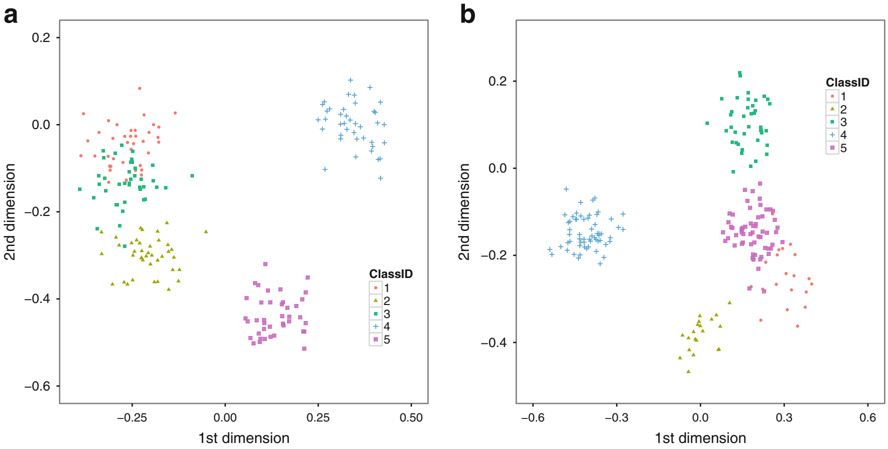


Fig. 1 Scatter plots of data points projected to the first two discriminant directions based on the true model. We set $n = 200$, $K = 5$, $q = 200$, $q_0 = 20$, $r = 2$, $h = 2$, $\sigma_1 = 1$, and $\sigma_2 = 0.05$. The *left panel* (**a**) is for balanced data with 40 samples in each class. The *right panel* (**b**) is for unbalanced data with 20, 20, 40, 60 and 60 samples in the five classes, respectively

Table 1 Simulation results: unbalanced data scenario

			ORE	LDA	PSFAR			SMR		
K	q	(σ_1, σ_2)	MCR	MCR	MCR	TPR	FPR	MCR	TPR	FPR
5	200	(1.0, 0.05)	5.1 (5.5)	53.4 (7.7)	7.9 (8.7)	92.2 (16.9)	7.3 (6.6)	13.4 (7.9)	42.6 (6.8)	10.2 (2.7)
		(0.6, 0.03)	2.2 (4.6)	35.3 (8.6)	2.3 (4.9)	97.0 (8.1)	9.1 (5.2)	6.5 (8.8)	38.8 (6.0)	6.7 (3.9)
	400	(1.0, 0.05)	5.9 (6.4)	23.4 (8.0)	6.5 (7.0)	97.9 (10.7)	9.1 (7.9)	15.6 (8.4)	36.0 (6.1)	5.0 (1.6)
		(0.6, 0.03)	1.6 (3.2)	10 (7.9)	1.7 (3.8)	98.3 (8.5)	14.0 (6.0)	7.8 (9.3)	33.1 (5.5)	3.1 (1.9)
10	200	(0.5, 0.05)	22.8 (5.6)	44.1 (5.3)	23.5 (5.7)	100.0 (0.0)	5.5 (5.6)	41.9 (4.9)	54.5 (3.9)	24.3 (1.5)
		(0.3, 0.03)	9.1 (5.2)	22.8 (6.6)	9.7 (5.4)	98.4 (5.3)	6.2 (5.4)	28.7 (6.3)	50.7 (4.5)	20.3 (1.7)
	400	(0.5, 0.05)	26.9 (2.8)	79.2 (1.0)	27.6 (3.4)	100.0 (0.0)	5.3 (5.6)	50.2 (5.0)	47.3 (2.9)	14.4 (0.8)
		(0.3, 0.03)	9.8 (6.1)	69 (4.6)	10.2 (6.1)	98.2 (5.7)	6.7 (6.3)	32.7 (6.8)	45.3 (4.4)	12.4 (0.8)

Reported are the misclassification rates (MCR) for classification, true positive rates (TPR), and false positive rates (FPR) for variable selection, with standard deviations in parenthesis

5 Classification of Images of Handwritten Digits

We use the MNIST handwritten digits data (LeCun et al. 1998) to test the performance gain of using PSFAR. In this dataset, the data from each image are the grey levels of $q = 28 \times 28 = 784$ pixels, and there are thousands of images for each digit. We use a random selection procedure to test the classification error rates of LDA and PSFAR. At each time, the training data is created by randomly selecting 50 samples from each digit "0" to "9", so that $n = 500$ and $K = 10$; the test data is created in the same way. The methods are applied on the training data and their classification performances are then evaluated using the test data. This procedure is repeated 50 times. The average MCR are 49.1 % (3.0 %) and 20.3 % (2.0 %) for LDA and PSFAR, respectively. Therefore, the performance of PSFAR is much better than that of LDA. We have also tested SMR, and its average MCR is 19.3 % (1.9 %). The reason that PSFAR does not outperform SMR is due to the fact that the data may severely deviate from the normality. Nevertheless, the dimensional reduction and joint mean and covariance estimation still allow PSFAR to be very competitive in this application.

Table 2 shows the proportions of time each digit is being classified to the ten possible digits, using PSFAR. For example, the last row shows that among all the digit "9"s appeared in the test datasets, 78.4 % of them are correctly classified while 10.0 % of them are classified to be "4". Figure 2 shows some images that are misclassified. It is clear that these hand-writing patterns are indeed quite different from the typical patterns of the digits.

Table 2 MNIST data: proportions of time each digit from test data is being classified to the ten possible digits using PSFAR. The proportions of correct classification are shown in bold

True\ Pred	0	1	2	3	4	5	6	7	8	9
0	**83.5**	0.2	1.3	1.7	0.9	4.2	2.5	0.9	4.1	0.7
1	0.0	**89.2**	0.9	0.4	0.3	0.9	0.2	0.2	7.9	0.1
2	0.9	5.3	**76.0**	2.8	1.4	1.1	4.5	2.5	4.3	1.3
3	0.7	1.9	3.9	**74.3**	0.4	8.4	0.7	2.7	5.1	1.9
4	0.1	1.1	0.7	0.2	**78.7**	0.7	2.2	0.7	2.5	13.0
5	1.5	1.3	0.7	7.1	2.3	**72.3**	1.5	1.0	10.3	2.3
6	1.3	1.6	0.7	0.1	1.9	2.3	**87.8**	0.2	4.1	0.1
7	0.6	3.3	0.7	0.9	1.3	0.4	0.1	**82.1**	1.8	8.7
8	0.5	7.0	1.2	4.9	1.1	6.9	1.7	1.2	**71.5**	3.9
9	0.4	0.4	0.3	1.5	10.0	1.7	0.3	4.5	2.6	**78.4**

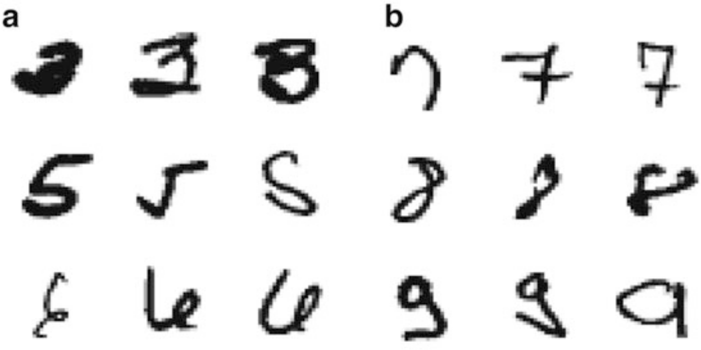


Fig. 2 MNIST data: misclassified images by PSFAR. The *left panel* (**a**) shows digits 3, 5, and 6 that are misclassified, from the *top to the bottom*. The *right panel* (**b**) shows 7, 8, and 9

References

Allwein, E. L., Schapire, R. E. and Singer, Y. (2001) Reducing multiclass to binary: A unifying approach for margin classifiers. *The Journal of Machine Learning Research*, **1**, 113–141.

Anderson, T. W. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, **22**, 327–351.

Bickel, P. J. and Levina, E. (2004) *Bernoulli*, **10**, 989–1010.

Bunea, F., She, Y. and Wegkamp, M. (2011) Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics*, **39**, 1282–1309.

Bunea, F., She, Y., Wegkamp, M. H. *et al.* (2012) Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, **40**, 2359–2388.

Cai, T. and Liu, W. (2011) A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, **106**, 1566–1577.

Chen, K., Chan, K.-S. and Stenseth, N. C. (2012) Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society Series B*, **74**, 203–221.

Chen, K., Dong, H. and Chan, K.-S. (2013) Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, **100**, 901–920.

Chen, L. and Huang, J. (2016) Sparse reduced-rank regression with covariance estimation. *Statistics and Computing*, **26**, 461–470.

Chen, L. and Huang, J. Z. (2012) Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, **107**, 1533–1545.

Fan, J. and Fan, Y. (2008) High dimensional classification using features annealed independence rules. *Annals of Statistics*, **36**, 2605.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Friedman, J., Hastie, T. and Tibshirani, R. (2001) *The elements of statistical learning*. Springer Series in Statistics Springer, Berlin.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1.

Glahn, H. R. (1968) Canonical correlation and its relationship to discriminant analysis and multiple regression. *Journal of the Atmospheric Sciences*, **25**, 23–31.

Gower, J. C. and Dijksterhuis, G. B. (2004) *Procrustes problems*. Oxford University Press.

Izenman, A. (2008) *Modern multivariate statistical techniques*. Springer.

Izenman, A. J. (1975) Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, **5**, 248–264.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**, 2278–2324.

Li, T., Zhu, S. and Ogihara, M. (2006) Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and Information Systems*, **10**, 453–472.

Lorena, A. C., Carvalho, A. C. P. L. F. and Gama, J. M. P. (2009) A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*, **30**, 19–37.

McLachlan, G. (2004) *Discriminant analysis and statistical pattern recognition*, vol. 544. John Wiley & Sons.

Rao, C. R. (1948) The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society Series B*, **10**, 159–203.

Reinsel, G. C. and Velu, P. (1998) *Multivariate reduced-rank regression: theory and applications*. New York: Springer.

Rothman, A. J., Levina, E. and Zhu, J. (2010) Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, **19**, 947–962.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011) Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.*, **39**, 1241–1265.

Vapnik, V. N. and Vapnik, V. (1998) *Statistical learning theory*. Wiley New York.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, **68**, 49–67.

Zhang, M.-L. and Zhou, Z.-H. (2014) A review on multi-label learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, **26**, 1819–1837.

Part VIII
Statistical Applications in Business and Finance

A Bivariate Random-Effects Copula Model for Length of Stay and Cost

Xiaoqin Tang, Zhehui Luo, and Joseph C. Gardiner

Abstract Copula models and random effect models are becoming increasingly popular for modeling dependencies or correlations between random variables. Recent applications appear in such fields as economics, finance, insurance, and survival analysis. We give a brief overview of the principles of construction of copula models from the Farlie-Gumbel-Morgenstern, Gaussian, and Archimedean families to the Frank, Clayton, and Gumbel families. We develop a flexible joint model for correlated errors modeled by copulas and incorporate a cluster level random effect to account for within-cluster correlations. In an empirical application our proposed approach attempts to capture the various dependence structures of hospital length of stay and cost (symmetric or asymmetric) in the copula function. It takes advantage of the relative ease in specifying the marginal distributions and introduction of within-cluster correlation based on the cluster level random effects.

Keywords Copula families • Random effects • Joint models • Healthcare cost

1 Introduction

Among the challenges in the analysis of multivariate outcomes of mixed types is the specification of a joint distribution that accommodates the different measurement scales and dependencies among the outcomes. In healthcare studies it is common to have multiple patient-level outcomes, some of which are continuous and others are discrete. For example, for hospital resource management and planning studies it is important to consider both length of stay (LOS) and the final disposition of

X. Tang
Asthma, Allergy and Autoimmunity Institute, Allegheny Health Network, 4800 Friendship Ave., Pittsburgh, PA 15224, USA
e-mail: atang@wpahs.org

Z. Luo • J.C. Gardiner (✉)
Department of Epidemiology and Biostatistics, Michigan State University, 909 Fee Road, B629 West Fee, East Lansing, MI 48824, USA
e-mail: zluo@epi.msu.edu; jgardiner@epi.msu.edu

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_25

the patient (died in hospital, discharged home, to nursing home, hospice, or other facility). During a hospital stay adverse events such as incident pressure ulcer, fall, or deep-vein thrombosis are regarded as defects in patient-care with the unintended consequences of an increase in LOS and cost. Flexible models that can address multiple outcomes of different types while incorporating covariates have useful application for prediction.

Multilevel models (also called hierarchical models, nested models, mixed models, random-effects (RE) models, random-coefficient models, or split-plot designs) are statistical models addressing variations at more than one level (Skrondal et al. 2004). They can be viewed as generalizations of linear models. For an example using LOS and cost, patients share characteristics of the hospital in which they are treated. Therefore, in addition to the patient-level dependence in outcomes (LOS, cost), there are dependencies between patients in the same hospital.

The multivariate normal (Gaussian) distribution is by far the most commonly used model for multivariate outcomes. However, multivariate normality might be an improper assumption in many situations. For example, if two outcomes are positive and skewed, the log-normal or log-logistic regression models might be more appropriate. This is especially true for LOS and cost where positive right skewness and correlation are present. In recent years, some attempts have been made to relate the two outcomes that permit consideration of the correlation between LOS and cost. Gardiner et al. (2002) propose a two-equation model for total cost and duration of treatment with the endogeneity of the latter accounted for in the model for cost. In their model correlation is assessed from a seemingly unrelated regression model for log-transformed cost and log-transformed duration. Although a bivariate normal distribution for these transformed outcomes is one consideration, better fit might be obtained by using for example, a log-normal distribution for cost and a log-logistic distribution for LOS. Other candidate distributions include the Pareto, Gamma and Weibull (Hossain et al. 2015; Gardiner et al. 2014).

The new approach that we consider in this article provides a more general and flexible model for the correlated variables with different distributions. Copula functions are useful tools to model dependence for multivariate outcomes. There is an increasing use of copulas in several scientific fields, such as economics (Trivedi et al. 2007), medical research (Lambert and Vandenhende 2002; Nikoloulopoulos and Karlis 2008) and finance and insurance (Bee 2004; Breymann et al. 2003; Klugman and Parsa 1999; Zhao and Zhou 2010). A copula is a function that connects the marginal distributions of the outcomes to restore the joint distribution with an association parameter which depends on the copula and not on the marginal distributions. For example, the multivariate Gaussian distribution can be generated by a Gaussian copula applied to Gaussian marginal distributions. The association parameter is a correlation matrix. More importantly, a copula can provide many flexible non-Gaussian joint distributions of varying complexity.

The context of our application of copula models is the bivariate outcome LOS and cost. They are likely to be correlated and have different marginal distributions. In addition we consider another potential correlation induced at the hospital (cluster) level from unmeasured latent variables (hospital efficiency, provider characteristics, etc.) that are shared by patients within the same cluster. By marginalization with respect to the distribution of the RE, we obtain the joint marginal distribution for the outcomes. However, this distribution typically does not have a closed-form.

RE models are used extensively in clustered and longitudinal data analysis to capture within-cluster dependencies. The linear mixed-effects model and generalized linear mixed-effects model are commonplace. Other applications of the RE model are in survival analysis where the term shared frailty is used for the RE that links the time-to-event outcomes, e.g, survival times of patients within the same cluster.

In this paper, we develop a new flexible joint model for outcomes based on correlated errors modeled by copulas and incorporate a cluster level RE to account for individual and within-cluster correlations simultaneously. The proposed approach captures the various dependence structures of LOS and cost (symmetric or asymmetric) in the copula function, and takes advantage of the relative ease in specifying the marginal distributions and introduction of within-cluster correlation based on the cluster-level RE. Our empirical application draws data on LOS and cost from the 2003 Nationwide Inpatient Sample (NIS) of the Healthcare Utilization Project (HCUP).

The paper is organized as follows. In Sect. 2 we give a brief outline to copulas in the context of bivariate models. Sect. 3 describes a bivariate copula for LOS and cost first without RE and then with RE in the context of our empirical application. A brief discussion concludes the paper.

2 Copulas and Dependence: Brief Overview

Let U_1, U_2 denote two possibly dependent random variables on the unit interval [0, 1]. A copula C defined on the unit square $(u_1, u_2) \in [0,1]^2 = [0, 1] \times [0, 1]$ is a continuous joint distribution function for (U_1, U_2) such that the marginal distributions are uniform on [0, 1]. Therefore for all $(u_1, u_2) \in [0,1]^2$, $C(u_1, 0) = 0, C(u_1, 1) = u_1, C(0, u_2) = 0, C(1, u_2) = u_2$. Three simple copulas are the *independence* copula Π, the *Fréchet lower bound* W and *upper bound* M given by

$$\Pi(u_1, u_2) = u_1 u_2, \;\; W(u_1, u_2) = \max\{0, u_1 + u_2 - 1\}, \;\; M(u_1, u_2) = \min\{u_1, u_2\}.$$

All copulas C are captured by the Fréchet bounds in the sense that $W \le C \le M$.

Let Y_1, Y_2 denote two random variables with marginal distributions F_1, F_2. Using a copula C we can construct a joint distribution $F(y_1, y_2) = C(F_1(y_1), F_2(y_2))$. Conversely, for a joint distribution F of (Y_1, Y_2) there exists a copula C for which this relationship holds. Moreover, C is unique if the marginals are continuous and $C(u_1, u_2) = F\left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right)$. With non-continuous marginals uniqueness obtains if we restrict C to the lattice of points on which (Y_1, Y_2) has positive probability. The practical use of a copuala is to link the marginals F_1, F_2 and thereby infuse dependence in (Y_1, Y_2). This dependence is a property of the copula and not of the marginals. When used in this manner a copula shares features with RE models where the joint distribution of (Y_1, Y_2) is constructed from marginals via RE.

Henceforth we will consider continuous joint distributions. The joint density f of (Y_1, Y_2) is $f(y_1, y_2) = c(F_1(y_1), F_2(y_2))f_1(y_1)f_2(y_2)$ where $c(u_1, u_2) = \frac{\partial^2 C(u_1,u_2)}{\partial u_1 \partial u_2}$ and f_1,f_2 are the marginal densities of (Y_1, Y_2).

A concordance coefficient for a copula C measures the extent to which the underlying variables (Y_1, Y_2) are rank-ordered, i.e., large values of one with large values of the other, or small values of one with small values of the other. Kendall's τ and Spearman's ρ are two measures of concordance. They depend only on the copulas and can be expressed as

$$\tau = 4E(C(U_1, U_2)) - 1 = 4\int_0^1\int_0^1 C(u_1, u_2)\, dC(u_1, u_2) - 1.$$

$$\begin{aligned}\rho &= 12E(U_1U_2) - 3 = 12\int_0^1\int_0^1 u_1u_2dC(u_1, u_2) - 3\\ &= 12\int_0^1\int_0^1 (C(u_1, u_2) - \Pi(u_1, u_2))\, du_1du_2.\end{aligned}$$

Both ρ and τ take values in [−1, 1]. Both measures are equal to zero for the independence copula Π, −1 for the Fréchet lower bound W, and +1 for the Fréchet upper bound M.

Tail dependence of the distribution $F(y_1, y_2) = C(F_1(y_1), F_2(y_2))$ is described by the *upper tail dependence measure* λ_u defined by

$$\lambda_u = \lim_{q\to 1-} P\left[Y_1 > F_1^{-1}(q) \middle| Y_2 > F_2^{-1}(q)\right] = 2 - \lim_{q\to 1-}(1 - C(q, q))/(1 - q),$$

and corresponding *lower tail dependence measure* λ_l

$$\lambda_l = \lim_{q\to 0+} P\left[Y_1 \leq F_1^{-1}(q) \middle| Y_2 \leq F_2^{-1}(q)\right] = \lim_{q\to 0+} C(q, q)/q.$$

Both λ_u, λ_l are zero for Π, 1 for W and zero for M.

A family of copulas $\{C_\theta : \theta \in \Theta\}$ is indexed by the *association parameter* θ. The concordance measures ρ_θ and τ_θ might have a more restricted range than [−1, 1] as θ varies over Θ. Because the parameters θ in different families have different ranges and interpretations, they are not comparable to each other. A *comprehensive family of copulas* is one that includes Π, W and M as members. Table 1 summarizes the concordance and tail dependency measures for selected copula families. For a thorough discussion of copulas see Nelson (2006).

Archimedean Copulas are defined by $C(u_1, u_2) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2))$ where the generator $\varphi : [0, 1] \to [0, \infty]$ is a continuous, convex, strictly decreasing function, with $\varphi(0) = \infty$, $\varphi(1) = 0$. Special forms of the generator φ define

Table 1 Summary of concordance and tail dependency measures for selected copula families

Copula	Spearman ρ_θ	Kendall τ_θ	Upper tail dependence $\lambda_{u,\theta}$	Lower tail dependence $\lambda_{l,\theta}$	Copula
Gaussian	$6\pi^{-1}\arcsin\left(\frac{1}{2}\theta\right)$	$2\pi^{-1}\arcsin(\theta)$	$[\theta = 1]$	$[\theta = 1]$	$C_\theta(u_1, u_2) = \Phi_2\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right), \theta \in [-1, 1]$
T	No closed form	$2\pi^{-1}\arcsin(\phi)$	$2T(-c, \nu + 1)^{\mathrm{a}}$	$2T(-c, \nu + 1)^{\mathrm{a}}$	$C_\theta(u_1, u_2) = T_2\left(T^{-1}(u_1), T^{-1}(u_2)\right),$ $\theta = (\nu, \phi),\ \nu \in (1, \infty),\ \phi \in [-1, 1]$
FGM	$\theta/3$	$2\theta/9$	0	0	$C_\theta(u_1, u_2) = u_1 u_2 (1 + \theta(1 - u_1)(1 - u_2)), \theta \in [-1, 1]$
Clayton	No closed form	$\theta/(\theta + 2)$	0	$2^{-1/\theta}$	$C_\theta(u_1, u_2) = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta},\ \theta \in (0, \infty)$
Frank	$1 - 12\theta^{-1}$ $(D_1(\theta) - D_2(\theta))$	$1 - 4\theta^{-1}$ $(1 - D_1(\theta))$	0	0	$C_\theta(u_1, u_2) = -\theta^{-1}\log\left(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_1} - 1)}{e^{-\theta} - 1}\right),$ $\theta \in (-\infty, 0) \cup (0, \infty)$
Gumbel-Hougaard	No closed form	$1 - \theta^{-1}$	$2 - 2^{1/\theta}$	0	$C_\theta(u_1, u_2) = \exp\left(-\left[\{-\log u_1\}^\theta + \{-\log u_2\}^\theta\right]^{1/\theta}\right),$ $\theta \in [1, \infty)$
Independence	0	0	0	0	$\Pi(u_1, u_2) = u_1 u_2$
Fréchet lower bound	-1	-1	0	0	$W(u_1, u_2) = \max\{0, u_1 + u_2 - 1\}$
Fréchet upper bound	1	1	1	1	$M(u_1, u_2) = \min\{u_1, u_2\}$

Φ is the standard normal distribution function, Φ_2 is the bivariate normal distribution function with correlation θ, unit variances, and zero means. T is the univariate central t-distribution function with ν degrees of freedom, T_2 is the bivariate t-distribution function with correlation ϕ and ν degrees of freedom

[a] $c = [(\nu + 1)(1 - \phi)/(1 + \phi)]^{1/2}$

FGM Farlie-Gumbel-Morgenstein

Debye functions $D_n(\theta) = n\theta^{-n}\int_0^\theta \frac{t^n}{(e^t - 1)}dt,\ n = 1, 2$

named copula families. Kendall's $\tau = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}dt$ but a simple expression for Spearman's ρ in terms of the generator is not available. The Archimedean copulas in Table 1 are the Clayton with generator $\varphi(t) = \theta^{-1}\left(t^{-\theta} - 1\right)$, the Frank with generator $\varphi(t) = -\log\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$ and the Gumbel-Hougaard with generator $\varphi(t) = (-\log t)^{\theta}$. There are several other copulas within and outside the Archimedian class. The Farlie-Gumbel-Morgenstern (FGM) copula is not Archimedian except for $\theta = 0$ in which case it is the independence copula Π with generator $\varphi(t) = (-\log t)$. A convex combination of copulas is a copula, and so are continuous mixtures of a familiy of copulas (Nelson 2006).

3 Example

The data set is a sample of discharges from 24 hospitals for the year 2003 in one mid-western state (Agency for Healthcare Research and Quality 2009). Patient-level covariates are: age at admission (restricted to 18–84 years), gender, race, a measure of overall presenting comorbidity as assessed by the Charlson Comorbidity Index (CCI) (Charlson et al. 1987; Pompei et al. 1991), and the number of procedures undergone (NPR). We also restrict to discharges with at least 1 day for LOS. The resulting data set has 12,152 discharges. Some characteristics of the sample are: female gender 61.7 %, white race 68.8 %, black race 14.1 %, age ≥65 years 39.8 %, no comorbidity (CCI = 0) 41.6 %, and no procedures (NPR = 0) 39.6 %. Each hospital discharge is associated with two utilization measures, LOS (in days) and total hospital charge (CHG in dollars).

3.1 Bivariate Model Without Random Effects

For analysis we start with separate regression models for the log-transformed outcomes $Y_1 = \log(LOS)$, $Y_2 = \log(CHG)$ given by $Y_{i1} = \mathbf{z}'_{i1}\beta_1 + \sigma_1\varepsilon_{i1}, Y_{i2} = \mathbf{z}'_{i2}\beta_2 + \sigma_2\varepsilon_{i2}$ where $\mathbf{z}_{i1}, \mathbf{z}_{i2}$ are the aforementioned covariates. Assuming log-logistic distributions for LOS and CHG is equivalent to assuming ε_{i1}, ε_{i2} have the logistic (survival) distribution $S(u) = (1 + e^u)^{-1}, -\infty < u < \infty$. Maximum likelihood estimation (MLE) provides estimates of the regression and scale parameters $(\beta_1, \beta_2, \sigma_1, \sigma_2)$.

Standardized residuals (SRES) and Cox-Snell residuals (CRES) are computed as: $s_{ik} = \left(Y_{ik} - \mathbf{z}'_{ik}\widehat{\beta}_k\right)/\widehat{\sigma}_k$ and $c_{ik} = -\log S\left(s_{ik}\right), k = 1, 2$ respectively. Under the assumed model $\{c_{ik} : 1 \le i \le n\}$ should behave like a sample from the exponential distribution with mean = 1 (Klein et al. 2003). Use proc LIFETEST in SAS software (SAS Institute Inc 2014) to estimate the cumulative hazard function H regarding CRES as "time". Overall fit can be gauged visually to see if there is gross departure from the exponential cumulative hazard $H_e(t) = t$ (Allison 2010). Figure 1 shows

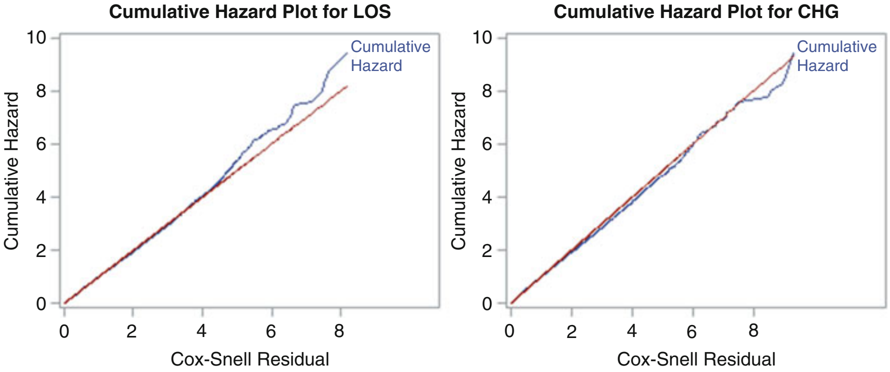


Fig. 1 Cumulative hazard plot of Cox-Snell residuals

Table 2 Copula association parameter estimates

Distribution	Parameter	Estimate	Standard error
T	DF, ν	8.1336	0.6758
	Correlation, ϕ	0.6149	...
Gaussian	Correlation, θ	0.6108	...
Clayton	Association, θ	0.8250	0.0166
Gumbel	Association, θ	1.7071	0.0123
Frank	Association, θ	4.5351	0.0640

the cumulative hazard plots for the CRES for LOS and CHG. A visual examination of the plots might suggest that the log-logistic model for CHG is acceptable, but for LOS it is quite tenuous. A more formal assessment of goodness of fit of parametric models for LOS could be made with Kolmogorov-Smirnov, Anderson-Darling or Cramer-von Mises statistics. For example, a Burr, Pareto or a Coxian phase-type might be appropriate for LOS (Tang et al. 2012; Gardiner 2012).

3.2 *Estimating a Copula Model*

We begin by assessing which of the five copulas in Table 1 would be a viable option for fitting a joint distribution to log-transformed (LOS, CHG). To address this objective the sample SRES $\{(s_{i1}, s_{i2}) : 1 \le i \le n\}$ is used as follows. First, from the empirical distribution functions (EDFs), $F_{1n}\left(y_1\right) = n^{-1}\sum_{i=1}^{n} [s_{i1} \le y_1]$, $F_{2n}\left(y_2\right) = n^{-1}\sum_{i=1}^{n} [s_{i2} \le y_2]$, the sample is transformed to pseudo data $\{(U_{i1}, U_{i2}) : 1 \le i \le n\}$ where the components have uniform marginals: $U_{i1} = F_{1n}\left(s_{i1}\right)$, $U_{i2} = F_{2n}\left(s_{i2}\right)$. Next, for each copula C the likelihood is constructed for the pseudo data and MLE gives estimates of the association parameters of the copula. The results of the five estimations are assembled in Table 2. The standard errors for the correlation in the T and Gaussian distributions are not shown because they are hardly used for inference.

Having estimated the association parameter we now simulate a sample of $B = 10{,}000$ draws from the copulas. The simulated sample $\{(\tilde{s}_{b1}, \tilde{s}_{b2}) : 1 \le b \le B\}$ has the same marginal distributions as the EDFs (F_{1n}, F_{2n}) of the original data. Figure 2 displays the results.

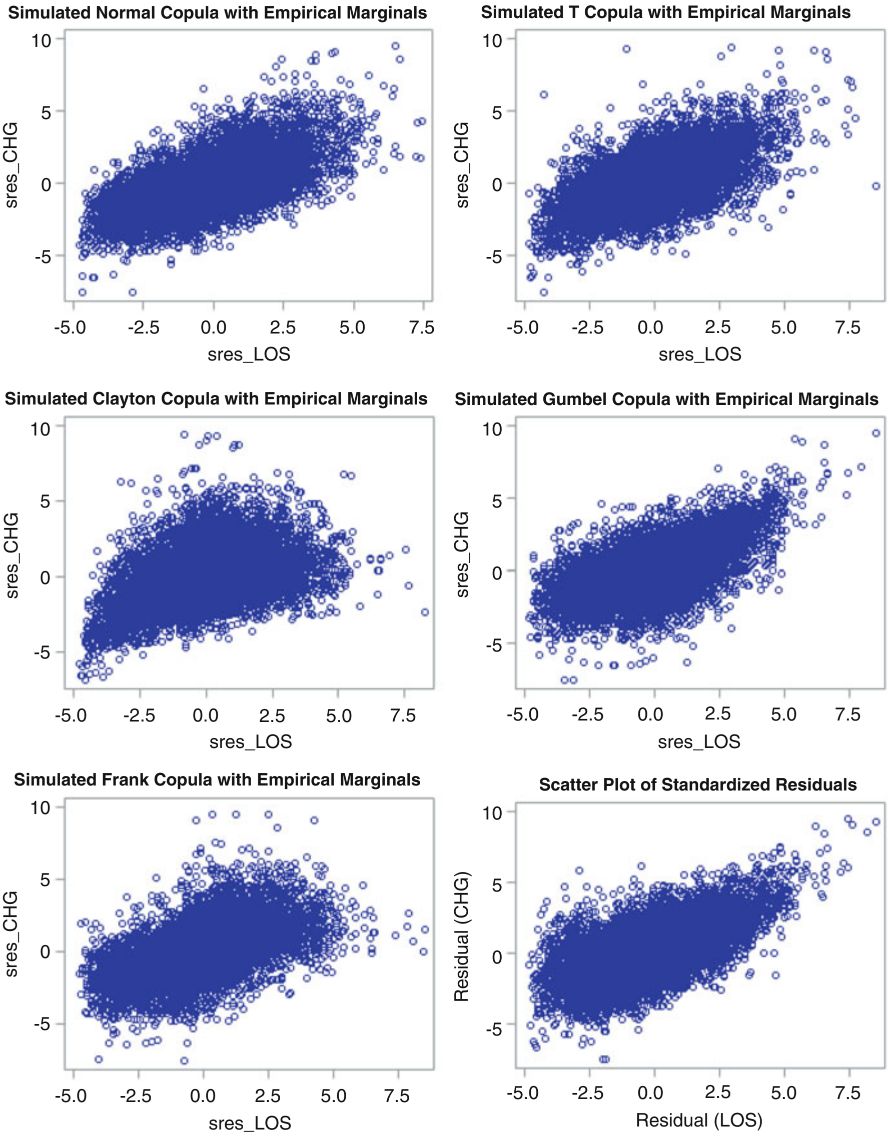


Fig. 2 Scatter plots of 10,000 simulated samples from five copulas and original data (*bottom right*)

The original scatter plot of the residuals ($n = 12{,}152$) is at the bottom right hand corner. Other scatter plots are from the simulated data ($B = 10{,}000$) of their respective copulas. Visual examination of these scatter plots suggests that the Gumbel-Hougaard copula is closer to the original data than any of the others. It also captures the upper-tail dependence. Comparisons based on Kolmogorov-Smirnov, Anderson-Darling or Cramer-von Mises statistics could be made (Kole et al. 2007).

3.3 Estimation of the Gumbel-Hougaard Copula

Our objective is to estimate the parameters of a bivariate Gumbel-Hougaard (GH) regression model for log-transformed (LOS, CHG), $Y_{i1} = \mathbf{z}'_{i1}\beta_1 + \sigma_1\varepsilon_{i1}, Y_{i2} = \mathbf{z}'_{i2}\beta_2 + \sigma_2\varepsilon_{i2}$ where ε_{i1}, ε_{i2} have marginal logistic distributions. From Table 1, the density function $c_\theta(u_1, u_2)$ of the copula is given by

$$c_\theta\left(u_1, u_2\right) = C_\theta\left(u_1, u_2\right)\left(u_1 u_2\right)^{-1}\left(\tilde{u}_1\tilde{u}_2\right)^{1-1/\theta}$$
$$\left(\left(\tilde{u}_1 + \tilde{u}_2\right)^{1/\theta} + \theta - 1\right) / \left(\left(\tilde{u}_1 + \tilde{u}_2\right)^{2-1/\theta}\right)$$

where $\tilde{u}_1 = \left(-\log u_1\right)^\theta$, $\tilde{u}_2 = \left(-\log u_2\right)^\theta$. Expressed in terms of $e_{i1} = \left(Y_{i1} - \mathbf{z}'_{i1}\beta_1\right)/\sigma_1$ and $e_{i2} = \left(Y_{i2} - \mathbf{z}'_{i2}\beta_2\right)/\sigma_2$, the joint density is $f\left(e_1, e_2\right) = c_\theta\left(F\left(e_1\right), F\left(e_2\right)\right) f\left(e_1\right) f\left(e_2\right)/\sigma_1\sigma_2$ where F and f are respectively, the standard logistic cumulative distribution and density functions. The previously fitted marginal distributions and assessment of the GH copula supply initial values for the model parameters $(\beta_1, \beta_2, \sigma_1, \sigma_2, \theta)$. Results of the MLE are in Table 3.

The estimates and their standard errors differ from their naïve counterparts from fitting marginal models (not shown), ignoring the association. If $\theta = 1$ the GH copula reduces to the independence copula. A formal test of $H_0 : \theta = 1$ would

Table 3 Gumbel-Hougaard copula for log(LOS) and log(CHG) with logistic marginals

		Log-logistic (LOS)			Log-logistic (CHG)		
Parameter	Class	Estimate	Std err	p-value	Estimate	Std err	p-value
Intercept		0.7814	0.0324	<0.0001	8.2306	0.0317	<0.0001
SEX (ref, male)	Female	−0.0066	0.0132	0.6180	−0.1832	0.0130	<0.0001
RACE (ref, white)	Black	0.1280	0.0181	<0.0001	0.0749	0.0181	<0.0001
	Other	0.0369	0.0166	0.0260	0.0376	0.0167	0.0246
AGE		0.0094	0.0004	<0.0001	0.0125	0.0004	<0.0001
CCI (ref, 3+)	0	−0.4309	0.0194	<0.0001	−0.3423	0.0190	<0.0001
	1	−0.2691	0.0193	<0.0001	−0.0989	0.0191	<0.0001
	2	−0.1461	0.0211	<0.0001	−0.0530	0.0208	0.0109
NPR (ref, none)	1+	0.3099	0.0130	<0.0001	0.8992	0.0128	<0.0001
Scale		0.3995	0.0029	<0.0001	0.3950	0.0029	<0.0001
Theta		1.7006	0.0145	<0.0001			

be rejected based on the Wald test, which is not surprising from the association seen in Fig. 2. Because testing H_0 places the parameter value on the boundary of the parameter space, the asymptotic distribution of the likelihood ratio test statistic is generally non-standard. A comparison of above model to a bivariate Gaussian copula model by a formal likelihood ratio test for two non-nested models (Vuong 1989) will support the GH copula. For additional discussion and SAS programs used in estimation see Gardiner (2013).

3.4 Bivariate Model with Random Effects

The model presented in Sect. 3.3 is modified to accommodate clustering effects of patients (j) within hospital (i): $Y_{ij1} = \mathbf{z}'_{ij1}\beta_1 + \nu_{i1} + \sigma_1\varepsilon_{ij1}$, $Y_{ij2} = \mathbf{z}'_{ij2}\beta_2 + \nu_{i2} + \sigma_2\varepsilon_{ij2}$. The addition of the two REs (ν_{i1}, ν_{i2}) incorporates correlation among patients within the same hospital (cluster) for each outcome, whereas $(\varepsilon_{ij1}, \varepsilon_{ij2})$ incorporates cross-equation correlation. All these random variables have zero means. Some assumptions are needed to avoid redunduncies and identify variance/covariance parameters. Assume $\{\varepsilon_{ijk} : 1 \le j \le n_i\}$ are independent and independent of (ν_{i1}, ν_{i2}). Then

$$Var\left(Y_{ijk}\right) = Var\left(\nu_{ik}\right) + \sigma_k^2 Var\left(\varepsilon_{ijk}\right) = \tau_k^2 + \sigma_k^2,$$

$$Cov\left(Y_{ijk}, Y_{ij'k}\right) = Var\left(\nu_{ik}\right) + \sigma_k^2 Cov\left(\varepsilon_{ijk}, \varepsilon_{ij'k}\right) = \tau_k^2,\ j \neq j'$$

$$Cov\left(Y_{ij1}, Y_{ij2}\right) = Cov\left(\nu_{i1}, \nu_{i2}\right) + \sigma_1\sigma_2 Cov\left(\varepsilon_{ij1}, \varepsilon_{ij2}\right) = \rho\tau_1\tau_2 + \theta\sigma_1\sigma_2$$

$$Cov\left(Y_{ij1}, Y_{ij'2}\right) = Cov\left(\nu_{i1}, \nu_{i2}\right) = \rho\tau_1\tau_2.$$

The intra-class correlation (ICC) between patients within the same hospital is $Corr\left(Y_{ijk}, Y_{ij'k}\right) = \tau_k^2 / \left(\tau_k^2 + \sigma_k^2\right)$, $k = 1, 2$, $j \neq j'$. The correlation for the two outcomes between patients is $Corr\left(Y_{ij1}, Y_{ij'2}\right) = \gamma\rho\tau_1\tau_2$, $j \neq j'$ and the correlation for the two outcomes within patients is $Corr\left(Y_{ij1}, Y_{ij2}\right) = \gamma\left(\rho\tau_1\tau_2 + \theta\sigma_1\sigma_2\right)$ where $\gamma = \left\{\left(\tau_1^2 + \sigma_1^2\right)\left(\tau_2^2 + \sigma_2^2\right)\right\}^{-1/2}$.

For each equation we proceed in a similar manner as before, estimating the regression and scale parameters $(\tau_k^2, \sigma_k^2, \beta_k)$. Estimates of the parameters ρ, θ are obtained through moments equations on the standardized residuals. These estimates are reasonable starting values of all parameters for the full MLE for a specified copula. For an application see Tang (2010). If all the random terms are assumed to have marginally normal distributions, the joint model for (Y_{ij1}, Y_{ij2}) becomes a linear mixed model. Estimation of parameters by either full MLE or restricted maximum likelihood (REML) is feasible with standard software (e.g., proc GLIMMIX in SAS software) (SAS Institute Inc 2014). In addition to the equation-specific regression parameters, two 2×2 covariance matrices are estimated for the variance

Table 4 Joint hierarchical model for log(LOS) and log(CHG)

Effect	Class	Lognormal (LOS)			Lognormal (CHG)		
		Estimate	Std err	p-value	Estimate	Std err	p-value
Intercept		0.8666	0.04151	<0.0001	8.2610	0.04955	<0.0001
SEX (ref, male)	Female	0.0099	0.01353	0.4668	−0.1598	0.01322	<0.0001
RACE (ref, white)	Black	0.0731	0.02372	0.0021	0.00033	0.02326	0.9886
	Other	0.0788	0.01980	<0.0001	0.04022	0.01946	0.0388
AGE		0.0079	0.00039	<0.0001	0.01144	0.00038	<0.0001
CCI (ref, 3+)	0	−0.4982	0.01967	<0.0001	−0.3844	0.01922	<0.0001
	1	−0.3462	0.01985	<0.0001	−0.1355	0.01939	<0.0001
	2	−0.1886	0.02165	<0.0001	−0.0601	0.02115	0.0045
NPR (ref, none)	1+	0.2641	0.01378	<0.0001	0.8418	0.01349	<0.0001

$\widehat{\tau}_1 = 0.1142,\ \widehat{\tau}_2 = 0.1780,\ \widehat{\rho} = 0.4325,\ \widehat{\sigma}_1 = 0.7038,\ \widehat{\sigma}_2 = 0.6876,\ \widehat{\theta} = 0.6272.$

components. It is recommended that the Cholesky parameterization be used. For example let $\begin{bmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{bmatrix} = \mathbf{CC}'$ where $\mathbf{C} = \begin{bmatrix} \alpha_1 & 0 \\ \alpha_{12} & \alpha_2 \end{bmatrix}$. Then $\tau_1^2 = \alpha_1^2,\ \rho\tau_1\tau_2 = \alpha_1\alpha_{12},\ \tau_2^2 = \alpha_{12}^2 + \alpha_2^2$. Estimation of the joint model can be realized through a single invocation to proc GLIMMIX applied to an expanded datset that has two records for each patient, one for each response type: RTYPE = 1 for log(LOS), RTYPE = 2 for log(CHG). Covariates effects are made specific to each response by crossing with RTYPE. Table 4 summarizes the results.

Chi-square tests show that ρ is not significantly different from one ($p = 0.06$), whereas θ is significantly different from one ($p < 0.0001$). Therefore, a slightly simpler model with $\rho = 0$ would suffice. Comparing the estimates and direction of covariate effects in Tables 3 and 4 shows that they are similar. Higher comorbidity, older age, and undergoing one or more procedures are associated with longer LOS and higher hospital charge. The effect of race seems different in the two models, but essentialy disappears when the model with $\rho = 0$ is estimated. A more thorough analysis with a richer constellation of covariates is not within the scope of the present article.

The fit of each model might be assessed by graphical techniques by plotting the standardized residuals against the quantiles of the assumed marginal distribution. This should be done separately for LOS and cost. In Fig. 3 the left hand panel is a quantile-quantile (QQ) plot from the RE model (Table 4); the right hand panel is the QQ-plot from the GH-copula model (Table 3).

For LOS both models indicate a short tail at the left end of the distribution. The GH copula model does better in the right end of the distribution of both LOS and charges (CHG). Note that the scales differ for the plots.

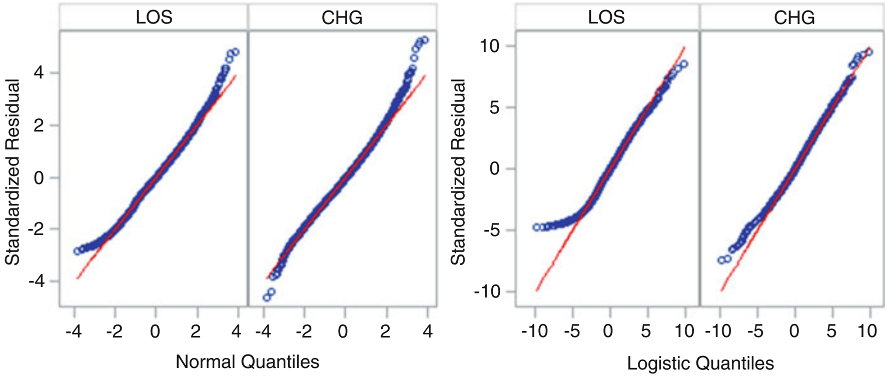


Fig. 3 QQ-plots of residuals from the random effects model (*left*) and the Gumbel-Hougaard copula model (*right*)

4 Discussion

We presented an application of a generalized mixed model and a copula model for joint analysis of bivariates outcomes. Although our empirical application has two continuous outcomes LOS and CHG, extension to multiple outcomes of mixed types is feasible. For example, Gardiner (2013) considers a trivariate model with LOS, CHG and a binary outcome of incident pressure ulcer during the hospital stay. See de Leon and Wu (2011) for a bivariate model where one component has multiple categories and the other component is a continuous variable. One purpose of copula modelling is to acknowledge the correlation between joint outcomes through the copula, while specifying the marginal distributions. It is generally much easier than positing a joint distribution. In estimation of parameters of the joint model, maximum likelihood is applicable in principle with initial values informed by the separately fitted marginal models. The association parameter of the copula must also be estimated and this can be done by considering candidate copula families and carrying out the estimation on simulated data that have the same marginals as the original data.

Although the theory of copulas has been in the literature for many decades, copula regression models (Kolev and Paiva 2009) have seen some interesting recent developments in empirical applications (Sacerdote and Sirovich 2010; Patton 2004; Shih 2014; Patton 2012). This growing field of research is gaining popularity in several areas, in economics, finance, insurance, and health services where correlated binary, count and continuous outcomes are dominant.

Acknowledgements The authors thank an anonymous reviewer for a careful review of the original manuscript and suggested changes for improvement. We also thank the organizers of the International Chinese Statistical Association (ICSA) for the invitation to present at the 2015 conference held in Fort Collins, CO.

References

Allison PD. *Survival Analysis using the SAS System–A Practical Guide. Second Edition.* Cary, NC: SAS Institute, Inc; 2010.

Bee M. Modelling credit default swap spreads by means of normal mixtures and copulas. *Applied Mathematical Finance.* 2004;11(2):125–146.

Breymann W, Dias A, Embrechts P. Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance.* 2003;3(1):1469–7688.

Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseseses.* 1987;40(5):373–383.

de Leon AR, Wu B. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine.* 2011;30(2):175–185.

Gardiner JC. Joint modeling of mixed outcomes in health services research, Paper 435–2013. Paper presented at: SAS Global Forum 2013; San Francisco, CA.

Gardiner JC. Modeling heavy-tailed distributions in healthcare utilization by parametric and Bayesian methods. Paper 418–2012. Paper presented at: SAS Global Forum 2012; Orlando, FL.

Gardiner JC, Luo Z, Bradley CJ, Polverejan E, Holmes-Rovner M, Rovner D. Longitudinal Assessment of Cost in Health Care Interventions. *Health Services and Outcomes Research Methodology.* 2002;3:149–168.

Gardiner JC, Luo Z, Tang X, Ramamoorthi RV. Fitting heavy-tailed distributions to health care data by parametric and bayesian methods. *Journal of Statistical Theory and Practice.* 2014;8(4):619–652.

HCUP Overview: Healthcare Cost and Utilization Project (HCUP). Rockville, MD: Agency for Healthcare Research and Quality; 2009.

Hossain MM, Laditka JN, Gardiner JC. The economic benefits of community health centers in lowering preventable hospitalizations: a cost-effectiveness analysis. *Health Services and Outcomes Research Methodology.* 2015;15(1):23–36.

Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data, 2nd Edition.* New York: Springer-Verlag; 2003.

Klugman SA, Parsa R. Fitting bivariate loss distributions with copulas. *Insurance: Mathematics and Economics.* 1999;24(1–2):139–148.

Kole E, Koedijk K, Verbeek M. Selecting copulas for risk management. *Journal of Banking & Finance.* 2007;31(8):2405–2423.

Kolev N, Paiva D. Copula-based regression models: A survey. *Journal of Statistical Planning and Inference.* 2009;139(11):3847–3856.

Lambert P, Vandenhende F. A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine.* 2002;21(21):3197–3217.

Nelson R. *An Introduction to Copulas, 2nd Edition.* New York, NY: Springer-Verlag; 2006.

Nikoloulopoulos AK, Karlis D. Multivariate logit copula model with an application to dental data. *Statistics in Medicine.* 2008;27(30):6393–6406.

Patton AJ. A review of copula models for economic time series. *Journal of Multivariate Analysis.* 2012;110:4–18.

Patton AJ. On the Out-of-Sample Importance of Skewness and Asymmetric Dependence for Asset Allocation. *Journal of Financial Econometrics.* 2004;2(1):130–168.

Pompei P, Charlson ME, Ales K, MacKenzie CR, Norton M. Relating patient characteristics at the time of admission to outcomes of hospitalization. *Journal of Clinical Epidemiology.* 1991;44(10):1063–1069.

Sacerdote L, Sirovich R. A copulas approach to neuronal networks models. *Journal of Physiology-Paris.* 2010;104(3–4):223–230.

SAS/STAT 13.2 User's Guide. Cary, NC: SAS Institute Inc; 2014.

Shih JH. Copula Models. In: Klein JP, vanHouwelingen HC, Ibrahim JG, Scheike TH, eds. *Handbook of Survival Analysis*. Boca Raton, FL: CRC Press; 2014.

Skrondal A, Rabe-Hesketh S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Boca Raton, FL: Chapman & Hall/CRC; 2004.

Tang X. *Modeling hospital length of stay and cost with heterogeneity, PhD dissertation*. East Lansing, MI: Statistics and Probability, Michigan State University; 2010.

Tang XQ, Luo ZH, Gardiner JC. Modeling hospital length of stay by Coxian phase-type regression with heterogeneity. *Statistics in Medicine.* 2012;31(14):1502–1516.

Trivedi PK, Zimmer DM. Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics, Vol 1*. Hanover, MA: NOW Publishers Inc; 2007.

Vuong QH. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica.* 1989;57(2):307–333.

Zhao XB, Zhou X. Applying copula models to individual claim loss reserving methods. *Insurance Mathematics & Economics.* 2010;46(2):290–299.

Index

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9