Springer Proceedings in Mathematics & Statistics

Mingxiu Hu · Yi Liu
Jianchang Lin *Editors*

Topics in Applied Statistics

2012 Symposium of the International Chinese Statistical Association

Springer

Springer Proceedings in Mathematics & Statistics

Volume 55

For further volumes:
http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Mingxiu Hu • Yi Liu • Jianchang Lin
Editors

Topics in Applied Statistics

2012 Symposium of the International Chinese Statistical Association

Springer

Editors
Mingxiu Hu
Takeda Pharmaceuticals
Cambridge, MA, USA

Yi Liu
Takeda Pharmaceuticals
Cambridge, MA, USA

Jianchang Lin
Takeda Pharmaceuticals
Cambridge, MA, USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-1-4614-7845-4 ISBN 978-1-4614-7846-1 (eBook)
DOI 10.1007/978-1-4614-7846-1
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013948449

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The 21st ICSA Applied Statistics Symposium was successfully held from June 23 to June 26, 2012, Boston, Massachusetts. International Society for Biopharmaceutical Statistics (ISBS) and American Statistical Association (ASA) co-sponsored the conference. There were 552 conference participants, 140 short course attendees, and 18 student volunteers. Over 400 papers were presented in 103 scientific sessions. Although the symposium has been held 21 times, this is the first time a proceeding is published for the symposium.

This inaugural proceeding of the ICSA Applied Statistics Symposium consists of 27 selected papers from a broad range of topics, from statistical applications in business and finance to applications in clinical trials and biomarker analysis. All papers have been reviewed by the editors to ensure quality but we do not expect the book is error-free. Instead of having a focused topic, we intentionally selected papers as broad as possible to serve the whole statistical community with all types of research interests. We specifically asked authors to explain all terminologies in simple forms so that statisticians from different research areas can understand and benefit from their papers. For example, papers in business should not be written with business statisticians in mind; instead, they should target the general population of statisticians. Therefore, readers of this book can enjoy not only papers from their own research areas but also those from different areas, which can potentially broaden their research interests.

In the part of *Statistical Applications in Business and Finance*, a house price index was constructed based on an autoregressive method that shows better predictive capabilities by accounting for changes in the market due to both single and repeat home sales. A method incorporating intra-day high-frequency trading data is demonstrated to result in better returns for stock portfolio allocation. A functional dynamic factor model is developed to forecast zero coupon bond yields for bond portfolio management and derivative security pricing. A mixed modeling approach is proposed for assessing the effectiveness of each marketing channel when multiple channels are used simultaneously in pharmaceutical promotions and

for optimizing the allocation of marketing resources. The uplifting modeling in identifying customers for targeted marketing campaigns through database analyses is discussed in the last paper of the section, and a tree-based uplifting modeling method is proposed for improving the traditional method in the evaluation of the marketing campaign efficiency.

In the *Biomarker Analysis and Personalized Medicine* section, the first paper points out that for some common efficacy measures such as relative risk or odds ratio, the efficacy for the overall population may not be representable as a function of efficacy for the subgroups and their prevalence unless certain design elements are satisfied. The second paper tries to identify subgroup of patients who benefit from beta-blocker treatment in a clinical trial setting in patients with chronic heart failure. The dataset is divided into two subsets, using the first set to build a parametric score for stratifying the remaining patients and the second set to obtain a nonparametric estimate of the treatment effect. Surrogate endpoints including biomarkers that can be obtained earlier in the trial to assess the efficacy of treatment are often of interest. The third paper considers a principal stratification approach to assess surrogacy for dichotomous markers and allows outcomes with missing values. A joint modeling of time-to-event clinical endpoint and longitudinal data on a biomarker is introduced in the fourth paper to assess correlation between the surrogate biomarker and the clinical endpoint. The last paper focuses on combining multiple continuous-scaled biomarkers into one single diagnostic or predictive tool for certain diseases.

In the *Bayesian Statistics in Clinical Trials* section, the authors illustrate how Bayesian methods are developed and applied in clinical studies. In early phase oncology trials, comparative simulation studies have demonstrated that the modified Toxicity Probability Interval (mTPI) design, a transparent Bayesian adaptive design, is safer in treating fewer patients at toxic doses above the MTD than the traditional 3+3 design. In another example of Bayesian application in early clinical studies, Bayesian inference is used for a biomarker interim analysis to facilitate the interim decision making for early clinical development. For the analysis of survival data from clinical trials, two new classes of semi-parametric survival models are proposed and demonstrated to have many practical advantages, including ease for determining priors, simpler interpretation of the regression parameters via the ratio of median survival times, and the ability to address heteroscedasticity of survival response. The last paper of this section presents a Bayesian conditional model for survival data with semi-competing risks in the presence of partial treatment switch, which is common due to ethical and practical reasons in clinical studies.

In the section of *High Dimensional Data Analysis and Statistical Genetics*, the first paper proposes a new pipeline with five steps for reverse-engineering the gene regulatory network using ordinary differential equation models. The second paper proposes a kernel-based semi-parametric regression method for testing the effect of microbiome composition on either a continuous or a binary outcome. A conditional autoregressive model is used to account for correlation in selecting

positive protein-coding genes to detect traces of natural selection. The fourth paper develops a sufficient dimension reduction method for high-dimensional regression with tensor predictors, which extends the conventional vector-based dimension reduction model. A successive standardization method is proposed and compared with the significance analysis of microarrays method in a case study in the last paper.

In the section of *Survival Analysis*, a synthesized approach is proposed that combines the observed overall survival (OS) effect and the predicted OS effect from the progression-free survival (PFS) data to explicitly test the implicit OS hypothesis at the time of primary analysis of PFS, which could help set an appropriate threshold for PFS effect size for accelerated approval from regulatory agencies. The other paper in this section considers a less restricted heterogeneity model to simultaneously analyze multiple cancer prognosis datasets. This model may provide an approach to address the reproducibility issue of findings about cancer prognosis markers across independent studies.

In the *Safety and Risk Analysis* section, the first paper reviews several approaches for analyzing low incidence adverse events with the consideration of drug exposure information and recurrence events in drug or vaccine development. The second paper tries to quantify through simulations the power in the setting of detecting cardiovascular signals in first-in-human trials and demonstrates that cardiovascular safety signals in general have reasonable statistical power for early detection when using a dose–response analysis.

In the *Longitudinal and Spatial Data Analysis* section, a new semi-nonparametric estimation approach is presented for constructing conditional reference charts (conditional quantiles) for muscular strength, a latent variable measured with error. The second paper develops a hierarchical Bayesian method for analyzing data with repeated binary responses over time and time-dependent missing covariates, which is a common problem in many biomedical studies and environmental applications.

In the *Multi-Regional Clinical Trials* section, a statistical method for multi-regional clinical trials is discussed for the setting where the treatment effect is similar but the variance is heterogeneous across different regions.

Finally, in the section of *In Vitro Drug Combination Studies*, an analytical method is proposed to evaluate different designs for in vitro drug combination studies, which typically involve a large number of wells with various concentrations of two drugs added together.

We are indebted to the authors who contributed their papers to the proceeding and carefully prepared their manuscripts in a tight timeline (2 months for the original draft and 1 month for incorporating our comments). Without a successful symposium, we would not have the source of quality papers to choose from for this book. Our deep gratitude goes to the Executive Organizing Committee (Mingxiu Hu, Tianxi Cai, Naitee Ting, Hongliang Shi, Minghui Chen, and Mark Chang), the Program Committee, and many other volunteers of the 21st ICSA Applied Statistics Symposium. We also appreciate Takeda Pharmaceuticals for giving us the time

to work on the book. We are grateful to Hannah Bracken and Marc Strauss of Springer for initiating the idea of publishing this proceeding and for their assistance throughout the entire process of completing the book.

Cambridge, MA, USA

Mingxiu Hu
Yi Liu
Jianchang Lin

Contents

Contributors

Gregory J. Anderson Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA

Lawrence D. Brown University of Pennsylvania, Philadelphia, PA USA

Davide Castagno University of Turin, Turin, Italy

Yuan-Chin Ivan Chang Institute of Statistical Science, Academia Sinica, Taipei Taiwan Department of Statistics, National ChengChi University, Taipei, Taiwan

Huaihou Chen New York University, New York, NY, USA

Jun Chen Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

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

Qingxia Chen Department of Biostatistics, Vanderbilt University, Nashville, TN, USA

Shein-Chung Chow Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA

Cong Chen Biostatistics and Research Decision Sciences, Merck Research Laboratories (MRL), Merck & Co., Inc, Upper Gwynedd, PA, USA

Brian Claggett Harvard School of Public Health, Boston, MA, USA

Michael R. Elliott Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA

Yu Fan Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, Houston, TX, USA

Lynda S. Gordon Merkle Inc., Columbia, MD, USA

Mike Hale Global Biostatistical Sciences, Amgen Inc., Thousand Oaks, CA, USA

Gregory Hather Takeda Pharmaceuticals, Cambridge, MA, USA

Spencer Hays Department of Statistical Sciences and Operations Research, Pacific Northwest National Laboratory, Virginia Commonwealth University, Richmond, VA, USA

Chin-Fu Hsiao Division of Clinical Trial Statistics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli, Taiwan

Jason C. Hsu The Ohio State University, Columbus, OH, USA

Eli Lilly and Company, Indianapolis, IN, USA

Man-Jen Hsu Department of Statistics, National ChengChi University, Taipei, Taiwan

Huey-Miin Hsueh Department of Statistics, National ChengChi University, Taipei, Taiwan

Peter H. Hu Bristol-Myers Suibb, Princeton, NJ, USA

Hui Huang Indiana University-Purdue University, Indianapolis, IN, USA

Jian Huang Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA, USA

Jianhua Z. Huang Department of Statistics, Texas A&M University, College Station, TX, USA

Jing Huang Global Biostatistical Sciences, Amgen Inc., Thousand Oaks, CA, USA

Lan Huang Office of Statistics, CDER, FDA, Silver Spring, MD, USA

Wei Huang Merkle Inc, Columbia, MD, USA

Joseph G. Ibrahim Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA

Yuan Ji NorthShore University HealthSystem Research Institute, Evanston, IL, USA

Eloise Kaizar The Ohio State University, Columbus, OH, USA

Lynn Kuo Department of Statistics, The University of Connecticut, Storrs, CT, USA

Paul O. Lewis Department of Ecology and Evolutionary Biology, The University of Connecticut, Storrs, CT, USA

Grace Ying Li Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN, USA

Hongzhe Li Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Yun Li Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA

Jianchang Lin Takeda Pharmaceuticals, Cambridge, MA, USA

Stuart Lipsitz Division of General Medicine, Brigham and Women's, Hospital, Boston, MA, USA

G. Frank Liu Merck Research Laboratories, Merck & Co. Inc, North Wales, PA, USA

Jin Liu Department of Biostatistics, School of Public Health, Yale University, New Haven, CT, USA

Ray Liu Millennium: The Takeda Oncology Company, Cambridge, MA, USA

Zhi-Ping Liu Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA

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

VA Cooperative Studies Program Coordinating Center, West Haven, CT, USA

Chaitra H. Nagaraja Fordham University, New York, NY, USA

Sang-Yun Oh Institute for Computational & Mathematical Engineering, Stanford University, Stanford, CA, USA

Richard A. Olshen Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA

Zhiying Pan Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA, USA

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

Xing Qiu Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA

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

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

Haipeng Shen Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Linda Z. Sun Biostatistics and Research Decision Sciences, Merck Research Laboratories (MRL), Upper Gwynedd, PA, USA

Te-Sheng Tan Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Miaoli County, Taiwan

Szu-Yu Tang The Ohio State University, Columbus, OH, USA

Jeremy M.G. Taylor Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA

Lu Tian Stanford University School of Medicine, Palo Alto, CA, USA

Pedro A. Torres Department of Mathematical Sciences, University of Puerto Rico at Mayagüez, Mayagüez, Puerto Rico

Michael T. Tsiang Department of Environmental Earth System Science, Stanford University, Stanford, CA, USA

Huixia Judy Wang Department of Statistics, North Carolina State University, Raleigh, NC, USA

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

Ouhong Wang Department of Biostatistics & Epidemiology, Amgen Inc, Thousand Oaks, CA, USA

Sue-Jane Wang Office of Biostatistics/Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA

Lee-Jen Wei Harvard School of Public Health, Boston, MA, USA

Hulin Wu Department of Biostatistics and Computational Biology, University of Rochester School of Medicine and Dentistry, Rochester, NY, USA

Rui Wu Department of Statistics, The University of Connecticut, Storrs, CT, USA

Shuang Wu Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA

Yuh-Jenn Wu Department of Applied Mathematics, Chung Yuan Christian University, Chung Li, Taiwan

Xiaodong Xue Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA, USA

Fang Yu Department of Biostatistics, University of Nebraska Medical Center, Omaha, NE, USA

Junjun Yue Merkle Inc, Columbia, MD, USA

Donglin Zeng Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA

Peng Zeng Department of Mathematics & Statistics, Auburn University, Auburn, AL, USA

Daowen Zhang Department of Statistics, North Carolina State University, Raleigh, NC, USA

Yuanye Zhang Novartis Institutes for BioMedical Research, Inc, University of Connecticut, Cambridge, MA, USA

Lihui Zhao Northwestern University School of Medicine, Chicago, IL, USA

Wenxuan Zhong Department of Statistics, University of Illinois at Urbana Champaign, Champaign, IL, USA

Jian Zou Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA

Part I
Applications in Business and Finance

Constructing and Evaluating an Autoregressive House Price Index

Chaitra H. Nagaraja and Lawrence D. Brown

Abstract We examine house price indices, focusing on an S&P/Case-Shiller-based index and an autoregressive method. Issues including the effect of gap time on sales and the use of hedonic information are addressed. Furthermore, predictive ability is incorporated as a quantitative metric into the analysis using data from home sales in Columbus, Ohio. When comparing the two indices, the autoregressive method is found to have the best predictive capabilities while accounting for changes in the market due to both single and repeat sales homes in a statistical model.

1 Introduction

Some uses for house price indices are as market indicators, prediction of individual house prices, or as inputs into other economic measures. However, houses are heterogenous goods varying in characteristics, age, and location. Furthermore, in any given time period, only a small subset of the population of homes is sold. Therefore, the price of a house is observed only intermittently. These are just some of the factors one must consider when constructing a house price index.

In essence, to create an index for a housing market, a geographic area (e.g., state, metropolitan area) and a time period (e.g., month, year) are selected. Sales data is collected and a summary measure is computed for each time period ($t = 0, \ldots, T-1$). The index is then constructed as a ratio of the measure at time period t to the base period measure ($t = 0$). Many types of models following this profile have

C.H. Nagaraja
Fordham University, 5 Columbus Circle, New York, NY 10019, USA
e-mail: cnagaraja@fordham.edu

L.D. Brown (✉)
University of Pennsylvania, 400 Jon M. Huntsman Hall,
3730 Walnut St., Philadelphia, PA 19104, USA
e-mail: lbrown@wharton.upenn.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_1,

been formulated including repeat sales indices which are the current benchmark index. In this paper, we describe a new autoregressive index method. This index allows for the use of both single and repeat sales in a statistical model with intuitive features automatically incorporated into the model. Furthermore, we provide the prediction of individual house prices as a method of index evaluation and show that the autoregressive method yields better predictions as well.

We begin in Sect. 2 with a description of the data which will be used to illustrate topics in the subsequent sections. Some background on the development of housing indices, in particular repeat sales indices, is provided in Sect. 3 followed by a description of the autoregressive index in Sect. 4. We compare indices in Sect. 5 and conclude in Sect. 6.

2 Data

The sales data used in this paper is part of a larger data set covering 20 metropolitan areas (see [16] and [17] for more details). Here, we focus on Columbus, Ohio in the USA as additional characteristics about the houses are available for sales in this area.

The data consists of single-family home sales which were approved for federally supported mortgages from July 1985 through September 2004. We divide the data into 3-month periods, resulting in 77 quarters, in order to have a sufficient number of sales per time period. A total of 162,716 sales of 109,388 unique homes were made in the sample period. Among these sales are 67,926 homes which are sold only once during this period. We have the time of sale, address (including ZIP (postal) code), and price for each sale. In addition, for 91% of the homes, we know the construction year, building area in square feet, number of bedrooms, and number of bathrooms; however, any structural changes in a house are not recorded. In the housing literature, these types of housing characteristics are called hedonic variables. Graphs of each of the hedonic characteristics, for the subset of data including those variables, are shown in Figs. 1 and 2.

To fit the models described in Sects. 3 and 4, the data was split into training and test sets. As will be described in Sect. 3, the repeat sales method can only be fit using homes that sell at least twice. Consequently, in order to fairly compare this method with the autoregressive approach, the test set must contain only final sales. The last sale from homes with three or more sales and the second sale of a random sample of sales from homes that sold twice were selected to create the test set; the remaining sales form the training set. Using this procedure, there were 136,989 sales in the training test and 25,727 sales in the test set. For the S&P/Case-Shiller® method (described further in Sect. 3), the training set observations resulted in 27,601 pairs of consecutive sales.

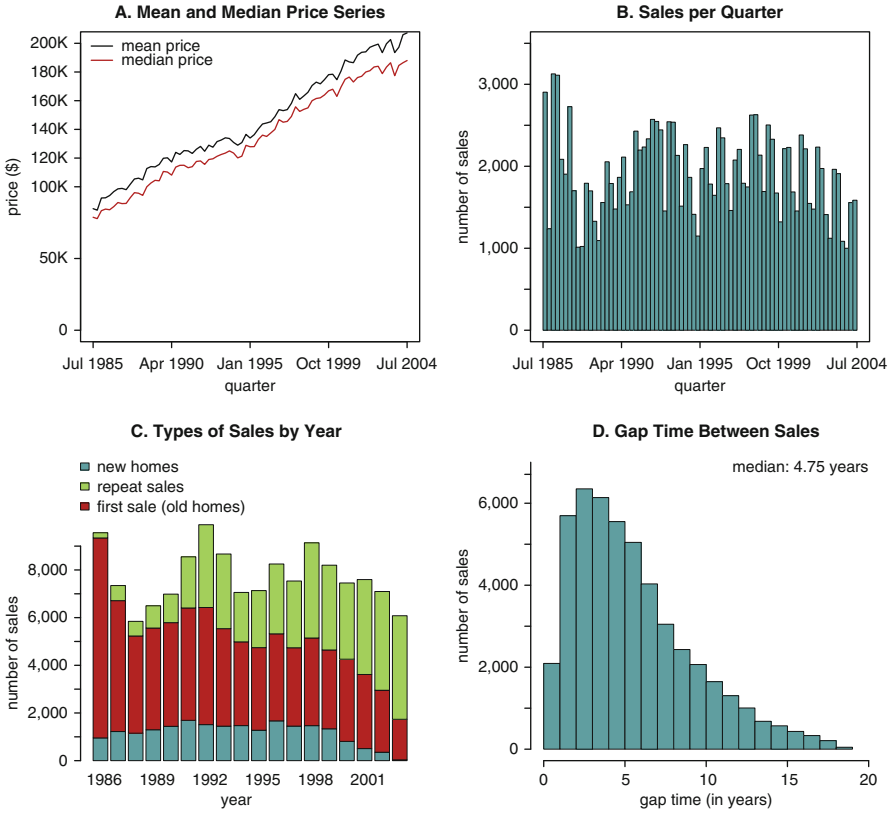


Fig. 1 Price series and histograms on home sales in Columbus, Ohio (July 1985–Sept. 2004).

3 Index Construction Methods

A basic index can be constructed from a median or mean price series where at each time period, the median or mean price is computed for a specified geographic area. In Fig. 1A, both the mean and median price series for Columbus, Ohio are plotted. Like income, house prices are heavily skewed as a few homes are very expensive. This is clear from Fig. 1A as the mean price series is consistently higher than the median price series. Therefore, many proposed indices use log prices instead of prices. The autoregressive method outlined in Sect. 4 uses log prices as well.

However, a median or mean price index is considered an insufficient measure of housing prices as it suffers from composition problems: (a) seasonal effects, (b) no control over types of houses sold each period, and (c) no quality adjustment. In Fig. 1B, the number of sales per quarter is graphed. We can see that during the winter, sales are much lower than in the summer. One contributing factor to this seasonal effect is that households with children tend to prefer to move in the summer

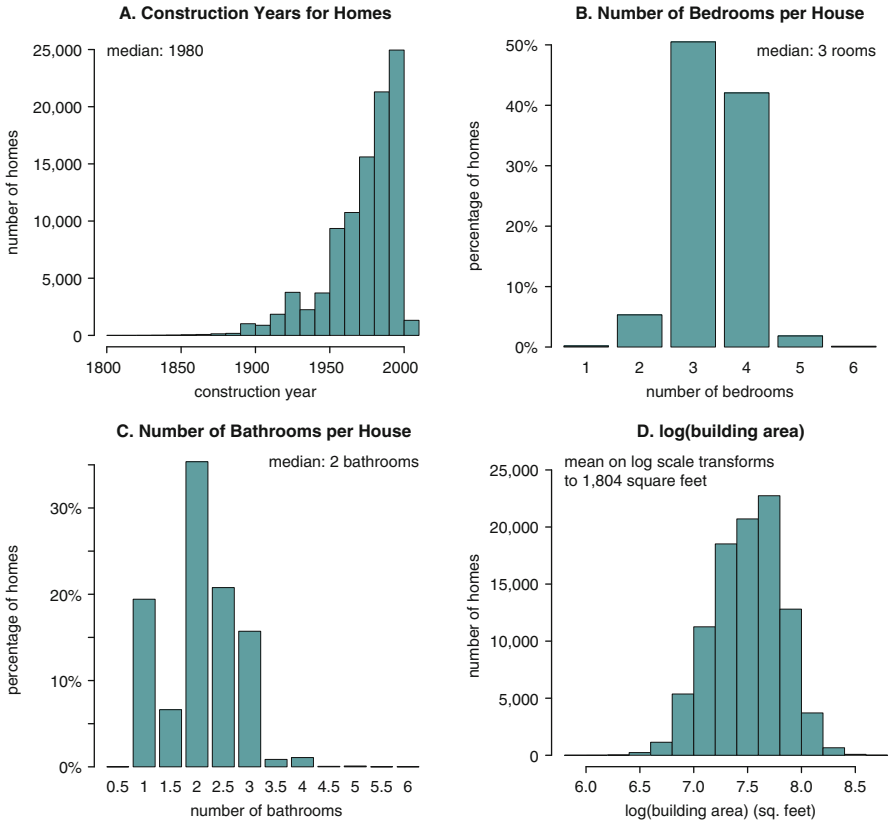


Fig. 2 Histograms and bar graphs for hedonic variables for homes in Columbus, Ohio: construction year, number of bedrooms and bathrooms, and building area in square feet.

and not during the school year. In Fig. 1C, the type of sales by year is graphed. Each bar represents a year and the height of the bar specifies the number of sales. The bars are then divided into three components illustrating the number of new homes (blue), old homes which are being sold for the first time in the data period (red), and old homes which have been sold more than once in the data period (green). As we will discuss next, new and old homes have quality differences which are ignored in median and mean indices. As a result, changes in a median or mean price index are not necessarily indicative of changes in the underlying market. While the National Association for Realtors publishes a median price series (Existing Home Sales Series) at the metropolitan area level, other indices have been developed to address these deficiencies.

Attributes of a house such as building area, number of bedrooms, and location are called hedonic characteristics. In their simplest form, hedonic characteristics can be used as predictor variables, along with time as a dummy variable, in a regression on price. The estimated time effects can be converted into an index by dividing

each by the base period effect. The resulting hedonic price index takes into account the heterogeneity of homes. In the Columbus, Ohio data, the available hedonic variables, construction year, number of bedrooms and bathrooms, and building area are plotted for unique homes in Fig. 2. We can see that homes in this city are quite varied. Many researchers advocate the use of hedonic indices [14]. However, given the difficulty of obtaining sufficient hedonic information and producing a robust model form, the hedonic index is considered impractical for general use.

The next development in house price index construction was the repeat sales index introduced by Bailey, Muth, and Nourse in 1963 [1]. This method divides the data into sets of sale pairs. The basic setup is as follows:

$$\begin{aligned} \frac{y_{it'}}{y_{it}} &= \frac{b_{t'}}{b_t} \times \varepsilon_{itt'} \\ \log y_{it'} - \log y_{it} &= \log b_{t'} - \log b_t + \log \varepsilon_{itt'} \end{aligned} \tag{1}$$

where y_{it} is the sale price of the ith house at time t, b_t is the price index, $t' > t$ for time periods $0, 1, \ldots, (T-1)$, and $\log \varepsilon_{tt'} \overset{iid}{\sim} N\left(0, \sigma^2\right)$ where *iid* denotes independent and identically distributed. The model as formulated in (1) is fit using least squares regression and by assumption, $b_0 = 1$ ($\log b_0 = 0$). There are two clear advantages of this model: the previous sale price is considered a proxy for the hedonic characteristics and the bias due to new homes, which are generally more expensive and are of higher quality, is eliminated [4].

Extensions by Case and Shiller [6, 7] and Calhoun [2], used in the Federal Housing Finance Agency House Price Index, and Shiller [21], implemented in the S&P/Case-Shiller® Home Price Indices, have different error structures than the original Bailey, Muth, and Nourse model. In the original method, the error term $\log \varepsilon_{itt'}$ is assumed to have the same variance regardless of the gap time between sales. (See Fig. 1D for a histogram of gap lengths.) However, Case and Shiller [6, 7] were the first to suggest that lengthy gaps between sales should be compensated for in the model. They do so by adding a random walk component to the error term. Consequently, $\log \varepsilon_{it} = \sum_{j=t^\star}^{t} v_{ij} + \eta_{it}$ where $v_{ij} \overset{iid}{\sim} N\left(0, \sigma_v^2\right)$ are steps of a Gaussian random walk, $t^\star$ is the time period when the house was constructed, and $\eta_{it} \overset{iid}{\sim} N\left(0, \sigma_\eta^2\right)$. Then, the error term in (1) is:

$$\begin{aligned} \log \varepsilon_{itt'} &= \log \varepsilon_{it'} - \log \varepsilon_{it} \\ &= \sum_{j=t+1}^{t'} v_{ij} + \eta_{it'} - \eta_{it} \end{aligned} \tag{2}$$

$$Var[\log \varepsilon_{itt'}] = (t' - t)\sigma_v^2 + 2\sigma_\eta^2. \tag{3}$$

As the gap time $(t' - t)$ increases, the variance of the error term also increases. The model is then fit using weighted least squares with estimates of the variances in (3) used as weights. Calhoun [2] proposes a similar extension as well. The S&P/Case-Shiller® index is estimated on the price scale instead of the log price scale; however,

the error structure is kept as in (2). This causes some problems in estimation which are managed to some extent by using instrumental variables [21, 22]. However, the disconnect between the model and error structure is difficult to support from a statistical perspective [15, 16].

There are numerous criticisms of repeat sales indices, two of which we discuss here. As repeat sales indices include only homes which have sold at least twice in the sample period, a large percentage of homes are excluded from the index [4, 14]. These single sales include both new homes and old homes sold only once during the sample period. Consequently, as can be seen in Fig. 1C, a large proportion of the data is excluded. In the overall data set, which covers over nearly 20 years, 42% of the sales are discarded because they are single sales. Omitting single sales homes can introduce bias into the index [9, 11–13]. One unintended consequence of discarding single sales is that two cities can have the same house price index even if one is growing at a much faster rate due to new construction than the other.

One of the prime assumptions of repeat sales indices is that hedonic characteristics are not required because all relevant information about the house is included in the previous sale price. In other words, (1) functions because the house at time t and at time t' are considered the same. There are two issues with this assumption. First, houses which have been significantly changed (remodeled or deteriorated) have to be excluded; with sufficient information about the house, this is possible. However, even if this is not the case, the house has aged over the intervening years between sales. Generally, price depreciates with age and is not accounted for in the repeat sales model [18]. As a remedy, adjustments incorporating age effects into the computed index have been proposed [3, 8, 10, 19]. Many of these modifications are added after the index calculations; therefore, we omit them in our analysis as they can always be added if implementing these indices.

Borrowing information across time periods as accomplished by the repeat sales indices is a valuable concept; however, excluding single sales is undesirable. A hybrid index, proposed by Case and Quigley [5], attempts to include single sales by combining repeat sales indices and hedonic index components. However, as with pure hedonic indices, such an index has high data requirements making it impractical for broad use. The autoregressive model described in Sect. 4 is a proposed index which includes both repeat and single sales in a suitable fashion without any additional data requirements in a statistical model.

4 Autoregressive Index

The autoregressive method as described in [17] considers all sales of the same house as components of a single series as opposed to pairs of sales. In theory, each house has a price at every time period. However, the price is observed only when sold. We start by defining the adjusted price x_{it} of the ith house at time of sale t ($t = 0, \ldots, T-1$).

$$x_{it} = \log y_{it} - \mu - a_t - \tau_z \tag{4}$$

where y_{it} is the sale price, μ is an overall mean for log prices, a_t is the log time effect in period t, and $\tau_z \overset{iid}{\sim} N\left(0, \sigma_\tau^2\right)$ is a random effect for ZIP (postal) code ($z = 1, \ldots, Z$), a proxy for location effects.

The adjusted prices follow a stationary AR(1) time series, observed when a sale occurs, and is a Markov process:

$$x_{it'} = \phi^{(t'-t)} x_{it} + \varepsilon_{it'}. \tag{5}$$

The parameter $|\phi| < 1$ is the correlation between adjusted prices of a house in consecutive time periods. The gap time between two sales, one at time t and the second at time t' where $t' > t$ is denoted as $(t' - t)$. For initial sales, there is no previous sale price and so the term $\phi^{(t'-t)} x_{it}$ drops out. To obtain a stationary series, the random variation term ε_{it} is defined as follows:

$$\begin{aligned} \varepsilon_{it} &\sim N\left(0,\ \sigma_\varepsilon^2 / \left(1 - \phi^2\right)\right) \quad \text{for the first sale} \\ \varepsilon_{it'} &\sim N\left(0,\ \sigma_\varepsilon^2 \left(1 - \phi^{2(t'-t)}\right) / \left(1 - \phi^2\right)\right) \quad \text{for subsequent sales} \end{aligned} \tag{6}$$

Where all of the error terms are independent. Two intuitive patterns arise naturally from this formulation of the model. First, as the time between sales increases, the correlation in adjusted prices between sales should decrease. The $\phi^{(t'-t)}$ component of (5) incorporates this feature. Second, as the time between sales increases, we also expect that the variance of the error term, ε_{it}, to increase as well which is visible through (6). The parameters $a_0, \ldots a_{T-1}, \sigma_\varepsilon^2, \sigma_\tau^2, \tau_1 \ldots \tau_Z$ are fit using the coordinate ascent procedure, an iterative maximum likelihood estimation method, as no closed-form estimates exist for most of the parameters. Further details can be found in [17]. One can also generalize the model in [17], by including additional hedonic information instead of only the ZIP code.

After fitting the data to the autoregressive method, a selection of the parameter estimates are: $\hat{\mu} = 11.5159$, $\hat{\phi} = 0.994807$, $\hat{\sigma}_\varepsilon^2 = 0.001264$, and $\hat{\sigma}_\tau^2 = 0.090329$. The value of $\hat{\mu}$ can be interpreted as an estimate of the overall average log price. Observe that the estimate for the autoregressive coefficient ϕ is very high, nearly one. One may argue that there is a unit root issue and that a random walk is a more appropriate model form (such as in the Case-Shiller model). However, this is not the case [17]. If we look back to Fig. 1D, the histogram of the gap time $(t' - t)$ between consecutive sales is graphed. The median gap time is 4.75 years or 14.25 quarters in Columbus, Ohio. Note that $\hat{\phi}^{14.25} = 0.928$ and is considerably less than one as a result. Given that nearly all sales occur with at least a gap of one year and most have a much longer gap time between sales, no appreciable unit root issue exists.

Finally, the log time effect a_t can be converted into a price index as follows: $b_0 = \exp\{\hat{a}_0 - \hat{a}_0\}, b_1 = \exp\{\hat{a}_1 - \hat{a}_0\}, \ldots, b_{T-1} = \exp\{\hat{a}_{T-1} - \hat{a}_0\}$. The first quarter, July through September 1985, is set as the base period and consequently $b_0 = 1$.

5 Comparison of Indices

No objective standards exist for evaluating and comparing indices. One reason is that there is no true index to compare an estimate to. Consequently, compiling evidence for or against a particular index is generally a matter of considering issues of sample selection bias based on the treatment of the data [4]. However, all of the indices described here are derived from models for the price of a house. Therefore prediction of individual prices can be used as an objective measure and as part of the index evaluation process. Given the limited availability of hedonic data, we do not examine the hedonic or hybrid indices here as comparisons to the repeat sales and autoregressive indices would not be fair. Instead, we focus on the S&P/Case-Shiller-based index, as it is the most well-known index, and the autoregressive index here. (A full set of comparisons with repeat sales indices can be found in [16].)

The model was fit using the training set observations (as described in Sect. 2) and predictions for repeat sales were computed on the test set observations. For the autoregressive method, to improve efficiency of the predictions when transforming predictions on the log scale back to the price scale, the following adjustment was made: $\hat{y}_{it} = \exp\{\widehat{\log y_{it}} + MSR/2\}$ where MSR is the mean squared error from the training set predictions [20]. As the S&P/Case-Shiller-based method is fit on the price scale, no similar adjustments are necessary. The test set root mean squared error, $RMSE = \left(\sum_{i=1}^{n}(y_{it} - \hat{y}_{it})^2\right)/n$, was compared between the two methods. For the autoregressive method, the test set RMSE = \$27,353 and for the S&P/Case-Shiller-based method, the RMSE = \$30,208. The autoregressive method has a lower RMSE for Columbus, Ohio indicating that this method produces better predictions. We find that the autoregressive method has the lowest RMSE values for all of the cities in the larger data set and amongst the other repeat sales methods [16, 17].

The last step is to compare index values. In Fig. 3, the mean (green), median (black), autoregressive (red), and S&P/Case-Shiller-based (blue) indices are plotted. We can see that the general trends are similar across indices. The mean and median indices fluctuate more because there is no smoothing over time periods. Unlike in Fig. 1A, when the mean and median price series differed, when converted into price indices, by dividing each value in the series by the base period price, they are not as different. Another feature is that the autoregressive index is nearly always between the mean price index and the S&P/Case-Shiller-based index. The mean price index treats all sales as if they were single sales; on the other hand, the S&P/Case-Shiller-based index incorporates only repeat sales observations. The autoregressive index includes both single sales and extracts the additional information from comparing repeat sales prices over time resulting in roughly a weighted average of the two index construction concepts.

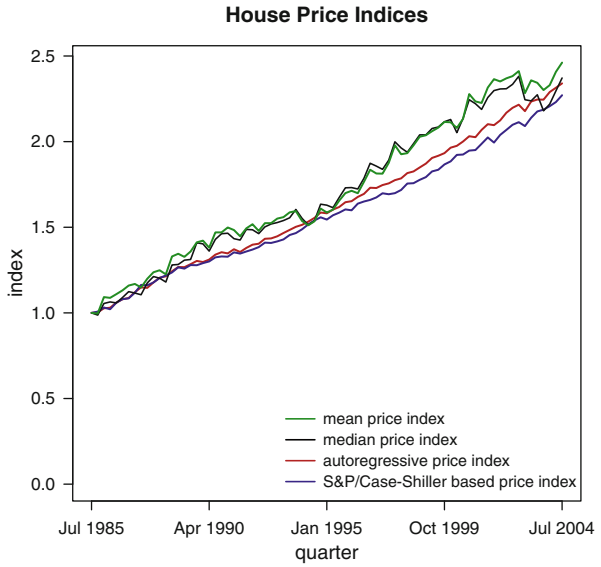


Fig. 3 Price indices for Columbus, Ohio: mean, median, autoregressive method, and S&P/Case-Shiller-based price indices.

6 Discussion

House price indices are useful for a variety of purposes and allow one to measure market changes over time. The autoregressive method described in this paper is an example of such an index. It not only incorporates the repeat sales idea inherent in traditional repeat sales methods, but also takes into account the vital information included in single sales which are omitted from repeat sales methods. The effect of gap times follows naturally from the model form by decreasing the correlation between adjusted prices and increasing the variance of the error term as the gap time increases. Both features intuitively make sense as large gap times would indicate that the previous sale price is less valuable and predicting the future price from the previous sale price becomes more difficult.

As these indices are synthetic constructions, they are difficult to evaluate. Adding the prediction of individual house prices to the list of evaluation techniques allows for a more objective measure of comparison. The autoregressive method performs better than the repeat sales methods on this metric as well, with lower test set RMSE values. All of these indices can be improved by removing homes which have changed drastically between sales and adding adjustments for the effects of age. Other refinements include adding relevant economic indicators or hedonic variables similar to the hybrid index idea set forth by Case and Quigley [5]. However, the autoregressive index provides us a good base form for an index to which we can incorporate additional features.

References

[1] Bailey, M.J., Muth, R.F. and Nourse, H.O. (1963). A regression method for real estate price index construction. *J. Amer. Statist. Assoc.* **58** 933–942.

[2] Calhoun, C. (1996). OFHEO house price indices: HPI technical description. Available at http://www.fhfa.gov/webfiles/896/hpi_tech.pdf .

[3] Cannaday, R.E., Munneke, H.J. and Yang, T.T. (2005). A multivariate repeat-sales model for estimating house price indices. *J. Urban Econ.* **57** 320–342.

[4] Case, B., Pollakowski, H.O. andWachter, S.M. (1991). On choosing among house price index methodologies. *AREUEA J.* **19** 286–307.

[5] Case, B. and Quigley, J.M. (1991). The dynamics of real estate prices. *Rev. Econ. Statist.* **73** 50–58.

[6] Case, K.E. and Shiller, R.J. (1987). Prices of single-family homes since 1970: New indices for four cities. *N. Engl. Econ. Rev.* **Sept./Oct.** 45–56.

[7] Case, K.E. and Shiller, R.J. (1989). The efficiency of the market for single family homes. *Amer. Econ. Rev.* **79** 125–137.

[8] Chau, K.W., Wong, S.K. and Yiu, C.Y. (2005). Adjusting for non-linear age effects in the repeat sales index. *J. Real Estate Fin. Econ.* **31** 137–153.

[9] Clapp, J.M. and Giacotto, C. (1992). Estimating price indices for residential property: a comparison of repeat sales and assessed value methods. *J. Amer Statist. Assoc.* **87** 300–306.

[10] Clapp, J.M. and Giacotto, C. (1998). Residential hedonic models: a rational expectations approach to age effects. *J. Urban Econ.* **44** 415–437.

[11] Clapp, J.M., Giacotto and C., Tirtiroglu, D. (1991). Housing price indices based on all transactions compared to repeat subsamples. *Amer. Real Estate and Urban Econ. Assoc. J.* **19** 270–284.

[12] Englund, P., Quigley, J.M. and Redfearn, C.L. (1999). The choice of methodology for computing housing price indexes: comparisons of temporal aggregation and sample definition. *J. of Real Estate Fin. Econ.* **19** 91–112.

[13] Gatzlaff, D.H. and Haurin, D.R. (1997). Sample selection bias and repeat-sales index estimates. *J. Real Estate Fin. Econ.* **23** 5–30.

[14] Meese, R.A. and Wallace, N.E. (1997). The construction of residential housing price indices: a comparison of repeat-sales, hedonic-regression, and hybrid approaches. *J. of Real Estate Fin. Econ.* **14** 51–73.

[15] Meissner, C., and Satchell, S. (2007). A comparison of the Case-Shiller house price index methodology with the FT house price index methodology. Available from http://www.acadametrics.co.uk .

[16] Nagaraja, C.H., Brown, L.D. and Wachter, S.M. (2014). Repeat sales house price index methodology. J. Real Estate Lit. *Forthcoming*.

[17] Nagaraja, C.H., Brown, L.D. and Zhao, L.H. (2011). An autoregressive approach to house price modeling. Ann. of Appl. Stat. **5**, 124–149

[18] Palmquist, R.B. (1979). Hedonic price and depreciation indexes for residential housing: a comment. *J. Urban Econ.* **6** 267–271.

[19] Palmquist, R.B. (1982). Measuring environmental effects on property values without hedonic regression. *J. of Urban Econ.* **11** 333–347.

[20] Shen, H., Brown, L. D. and Zhi, H. (2006). Efficient estimation of log-normal means with application to pharmacokinetic data. Statist. Med. **25** 30233038.

[21] Shiller, R. (1991). Arithmetic repeat sales price estimators. *J. Housing Econ.* **1** 110–126.

[22] S&P/Case-Shiller® Home Price Indices: Index Methodology (November 2009). Available at http://www.standardandpoors.com .

On Portfolio Allocation: A Comparison of Using Low-Frequency and High-Frequency Financial Data

Jian Zou and Hui Huang

Abstract Portfolio allocation is one of the most fundamental problems in finance. The process of determining the optimal mix of assets to hold in the portfolio is a very important issue in risk management. It involves dividing an investment portfolio among different assets based on the volatilities of the asset returns. In the recent decades, it gains popularity to estimate volatilities of asset returns based on high-frequency data in financial economics. However there is always a debate on when and how do we gain from using high-frequency data in portfolio optimization. This paper starts with a review on portfolio allocation and high-frequency financial time series. Then we introduce a new methodology to carry out efficient asset allocations using regularization on estimated integrated volatility via intra-day high-frequency data. We illustrate the methodology by comparing the results of both low-frequency and high-frequency price data on stocks traded in New York Stock Exchange over a period of 209 days in 2010. The numerical results show that portfolios constructed using high-frequency approach generally perform well by pooling together the strengths of regularization and estimation from a risk management perspective.

1 Introduction

Portfolio allocation is one of the most fundamental problems in finance. It involves dividing an investment portfolio among different assets. The process of determining the optimal mix of assets to hold in the portfolio is a very important issue in risk management. Markowitz (1952) was the original milestone paper for modern portfolio theory on the mean-variance analysis by solving an unconstrained quadratic optimization problem. It was later expanded in the book Markowitz

J. Zou (✉) • H. Huang
Indiana University-Purdue University, Indianapolis, IN 46202, USA
e-mail: zoujian@iupui.edu; huanghui86@gmail.com

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_2,

(1959). This approach has had a profound impact on financial economics. Sharpe (1966) introduced the Sharpe ratio for the performance of mutual funds, which is a direct measure of reward-to-risk.

High-frequency financial data usually refer to intra-day observations. The extra amount of information contained in high-frequency data enables researchers to develop more accurate estimators for the volatility matrix. The volatility matrix of daily asset returns is a key input in portfolio allocation, option pricing, and risk management. However, the main challenge is that when the number of assets is large, the volatility matrix cannot be estimated accurately. To address this problem, several innovative approaches for volatility matrix estimation were proposed in the past decade. Estimation methods for the univariate case include realized volatility (RV) (Andersen et al. (2001)), bi-power realized variation (BPRV) (Barndorff-Nielsen and Shephard (2002)), two-time scale realized volatility (TSRV) (Zhang et al. (2005)), wavelet realized volatility (WRV) (Fan and Wang (2007)), kernel realized volatility (KRV) (Barndorff-Nielsen et al. (2008)), and Fourier realized volatility (FRV) (Mancino and Sanfelici (2008)). For multiple assets, we face non-synchronization issue, which is referred to as that high-frequency price data are not aligned properly for different assets, hence are recorded at various mismatched time points. For a pair of assets, Hayashi and Yoshida (2005) and Zhang et al. (2005) have developed methods based on overlap intervals and previous ticks to estimate co-integrated volatility of the two assets. Barndorff-Nielsen and Shephard (2004) have considered estimation of integrated co-volatility for synchronized high-frequency data.

From a financial risk management perspective, Jagannathan and Ma (2003) analyzed the impact of weights constraints in large portfolio allocation. They show that solving the global minimum variance portfolio problem with some constraints on weights is equivalent to use a shrinkage estimate of the covariance matrix. Fan et al. (2012a) studied portfolio allocation with gross-exposure constraint combining vast volatility matrix estimation using different sampling schemes. However, there exists an interesting question as of when and how an investor will benefit from using high-frequency financial data in his/her portfolio allocation decisions. This paper aims to perform a comparative study on the performance of portfolios constructed using both low-frequency and high-frequency financial data. Therefore, we hope to shed some light on the benefits of high-frequency data in risk management.

The data that motivated this research comprise the transaction-by-transaction stock prices from the 30 DJIT composite constituents traded in New York Stock Exchange (NYSE). The data set is huge with ultra high-frequency observations since these stocks are highly liquid with vast trading volume. These high-frequency financial data also possess unique features such as price discreteness, unequally spaced time intervals, and non-synchronized trading (see, e.g., Wang and Zou (2010) for some illustrations of these issues). The normal trading hours of the NYSE are from 09:30 until 16:00. Thus, for simplicity, we discarded any transactions beyond these hours from our analysis.

The rest of this paper is structured as follows. Section 2 elaborates the framework of our methodology of portfolio allocation for high-frequency data. Section 3 presents numerical evidence on the performance comparison of our method under both low- and high-frequency scenarios. Section 4 concludes this paper.

2 Methods

Suppose that a portfolio consists of p assets and their log price process $X_t = (X_{1t}, \cdots, X_{pt})^T$ obeys an Itô process governed by

$$dX_t = \mu_t\, dt + \sigma_t\, dW_t, \qquad t \in [0, L], \tag{1}$$

where W_t is a p-dimensional standard Brownian motion, μ_t is a drift taking values in R^p, and σ_t is a diffusion variance of $p \times p$ matrix. Both μ_t and σ_t are assumed to be continuous in t.

For the portfolio with allocation vector w (i.e., percentage of each asset in the portfolio) and a holding period T, the variance (risk) of the portfolio return is given by $R(w, \Sigma) = w^T \Sigma w$. However, it is well known that the estimation error in the mean vector μ_t could severely affect the portfolio weights and produce suboptimal portfolios. This motivates us to adopt another popular portfolio strategy: the *global minimum variance portfolio*, which is the minimum risk portfolio with weights that sum to one. These weights are usually estimated proportional to the inverse covariance matrix, i.e., $w \propto \Sigma^{-1}$. Following Jagannathan and Ma (2003) and Fan et al. (2012b), we consider the following risk optimization with two different constraints:

$$\min w^T \Sigma w, \qquad s.t.\, \|w\|_1 \leq c \text{ and } w^T \mathbf{1} = 1 \tag{2}$$

where c is the gross exposure parameter which specifies the total exposure allowed in the portfolio and $\|\cdot\|_1$ is the standard vector L_1 norm. The summation to one constraint ensures weight percentages sum to 100%, inducing a full investment. An additional common requirement is no short selling (e.g., 401k, IRAs, and mutual funds), which adds the nonnegative optimization constraint. This corresponds to the case $c = 1$ as in Jagannathan and Ma (2003). The second optimization case is the global minimum risk portfolio where the gross exposure constraint $c = \infty$. Note that we only consider these two cases for simplification of the problem. Other cases with varying c can be easily generalized in our methodology.

2.1 Average Realized Volatility Matrix for High-Frequency Financial Data

The realized co-volatility computed from the high-frequency data on the prices of the p assets is used to estimate the integrated volatility over a given period of time. Let's take a day as a unit time. The integrated volatility for the ℓ-th day is a matrix defined to be

$$\Sigma_x(\ell) = \int_{\ell-1}^{\ell} \sigma_s \sigma_s^{\dagger} ds, \qquad \ell = 1, \cdots, L. \tag{3}$$

We first consider estimation of $\Sigma_x(1)$ on day one. For $i = 1, \cdots, p$, suppose that the high-frequency data for the i-th asset on day one are observed at times t_{ij}, $j = 1, \cdots, n_i$, and we denote by $Y_i(t_{ij})$ the observed log price of the i-th asset at time t_{ij}. Because transactions may be made at distinct times for different assets, the high-frequency price data for p-assets are often recorded at various mismatched time points. This is referred to as non-synchronized problem in high-frequency financial data.

Another complication in high-frequency financial data is that due to microstructure noise, the observed log price $Y_i(t_{ij})$ of the i-th asset is a noisy version of its true log price $X_i(t_{ij})$ and is usually assumed to obey an additive noise model

$$Y_i(t_{ij}) = X_i(t_{ij}) + \varepsilon_i(t_{ij}), \tag{4}$$

where $\varepsilon_i(t_{ij})$ are assumed to be i.i.d. noise with mean zero and finite fourth moments and independent of $X_i(t)$.

We consider realized co-volatility based on previous-tick times. Let $n = n_1 + \cdots + n_p$ and assume that n_i/n are bounded away from zero. Take a predetermined sampling frequency $\tau_k, k = 1, \cdots, m$. One such example is to select τ_i be a regular grid. For each $k = 1, \cdots, m$, we choose the corresponding observation time for the i-th asset by

$$\tau_{ik} = \max\{t_{ij} \le \tau_k, j = 1, \cdots, n_i\}, \qquad i = 1 \cdots, p.$$

The realized co-volatility is defined to be a $p \times p$ matrix $\hat{\Sigma}_y(1)$ whose (i_1, i_2) element is given by

$$\hat{\Sigma}_y(1)[i_1, i_2] = \sum_{k=1}^{m} [Y_{i_1}(\tau_{i_1,k}) - Y_{i_1}(\tau_{i_1,k-1})]\,[Y_{i_2}(\tau_{i_2,k}) - Y_{i_2}(\tau_{i_2,k-1})].$$

We estimate $\Sigma_x(1)$ by $\hat{\Sigma}_y(1)$. Hayashi and Yoshida (2005) showed that when there is no microstructure noise and $m \sim n$, $m^{1/2}(\hat{\Sigma}_y(1) - \Sigma_x(1))$ has asymptotic mixed normal distribution. Their result implies that as $n \to \infty$,

$$\hat{\Sigma}_y(1) - \Sigma_x(1) = O_P(n^{-1/2}).$$

Zhang (2006) proved that with i.i.d. microstructure noise and $m \sim n^{2/3}$,

$$\hat{\Sigma}_y(1) - \Sigma_x(1) = O_P(n^{-1/3}).$$

We apply the realized co-volatility to the price observations on the ℓ-th day and obtain the realized co-volatility estimator $\hat{\Sigma}_y(\ell)$ of integrated volatility $\Sigma_x(\ell)$ on the ℓ-th day. Above asymptotics imply that

$$\hat{\Sigma}_y(\ell) - \Sigma_x(\ell) = O_P(n^{-\eta}), \qquad \ell = 1, \cdots, L, \tag{5}$$

where η is equal to $1/3$ for i.i.d. noise and $1/2$ for no noise in the price data. However, since there are other sources of randomness in the data, such as price discreteness and non-synchronization errors, Wang and Zou (2010) showed that the rate η is on the order of $1/6$. Subsequently, Tao et al. (2011) established the optimal minimax risk for estimating large volatility matrix under the subguassian tail assumption.

3 Numerical Studies

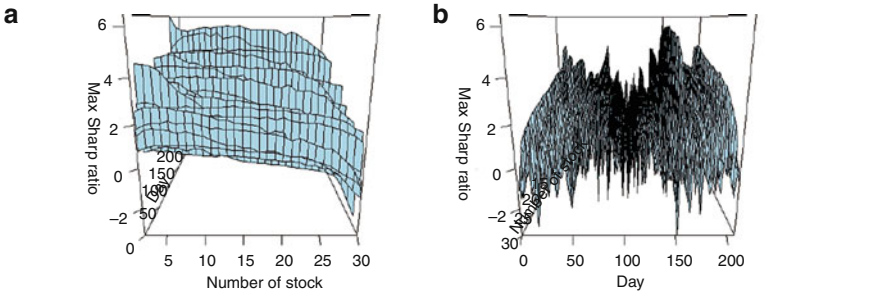
3.1 Portfolio Optimization with Low Frequency Data

First, we applied our methodology to the daily closing prices of Dow 30 stocks in 2011. The data are publicly available in many online search engines such as Yahoo Finance. Since the number of stocks we considered here is 30, we calculated the minimal variance portfolio consisting from 2 stocks to 30 stocks. The main challenge here is the possible portfolio combination increases exponentially when the number of stocks increases. If considering all the possibilities, the total case would be on the order of 2^{30}. This would be prohibitive to do in real applications. Hence we resort to sampling method to get around this issue. For a fixed number of stocks, we randomly generate 10,000 samples from the 30 stocks and then performed the calculation of yield and variance based on each subsample. We recorded the global minimal variance portfolio and the corresponding sharp ratio, which is defined as

$$S = \frac{E[R - R_f]}{\sigma} = \frac{E[R - R_f]}{\sqrt{\text{var}[R_a - R_b]}}. \tag{6}$$

with R_f being the risk free rate. It is a measure of the excess return (or risk premium) per unit of deviation in an investment asset, typically referred to as risk (and is a deviation risk measure).

In most cases, the maximum sharp ratio first increases when we include more stock but decreases when the number of stock exceeds certain threshold (Fig. 1a). For each day, the cusp point is different, ranging from two to seven visually. If we



Relationship between maximum sharp ratio and number of stocks

Dynamics of maximum sharp ratio across the whole year

c

Aug 17, 2011	11,392.01	11,529.67	11,322.30	11,410.21	171,281,254
Aug 16, 2011	11,480.48	11,488.01	11,292.63	11,405.93	187,802,137
Aug 15, 2011	11,269.85	11,484.60	11,269.85	11,482.90	188,122,365
Aug 12, 2011	11,143.46	11,346.67	11,142.18	11,269.02	228,034,779
Aug 1, 2011	10,729.85	11,278.90	10,729.85	11,143.31	393,188,013
Aug 10, 2011	11,228.00	11,228.00	10,686.49	10,719.94	396,303,407
Aug 9, 2011	10,810.91	11,244.01	10,604.07	11,239.77	431,414,648
Aug 8, 2011	11,433.93	11,434.09	10,809.85	10,809.85	479,981,128
Aug 5, 2011	11,383.98	11,555.41	11,139.00	11,444.61	406,307,400
Aug 4, 2011	11,893.86	11,893.94	11,372.14	11,383.68	300,764,953
Aug 3, 2011	11,863.74	11,904.91	11,700.34	11,896.44	198,216,482
Aug 2, 2011	12,129.77	12,130.30	11,865.56	11,866.62	207,057,660
Aug 1, 2011	12,144.22	12,282.42	11,998.08	12,132.49	182,822,673

The high volatility region in Figure 1b corresponding to the high volatility of the whole market

Fig. 1 Portfolio analysis using low-frequency data

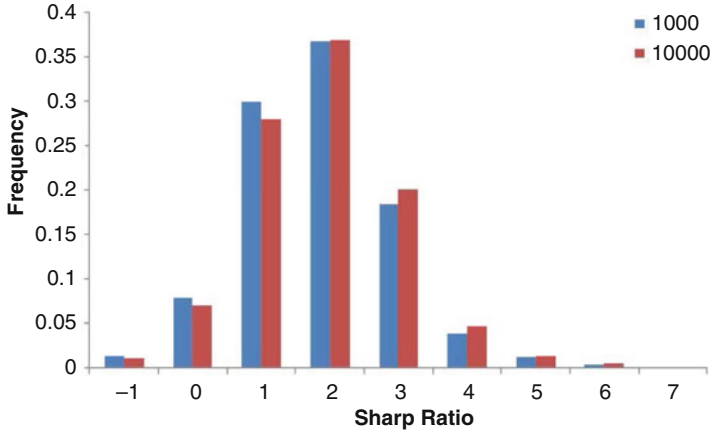
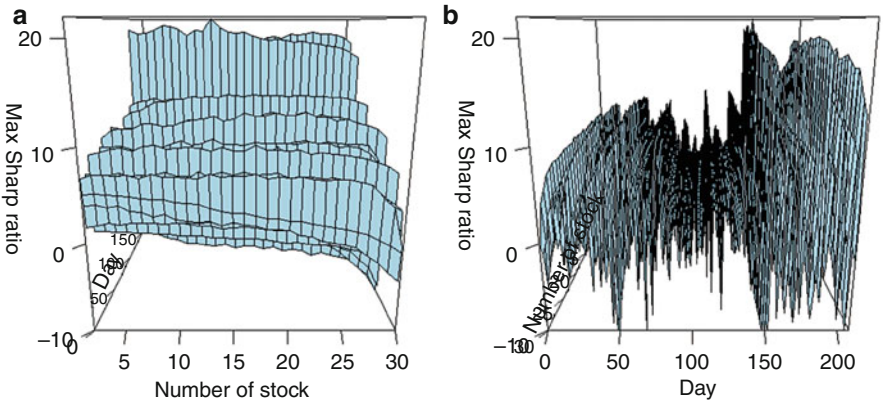


Fig. 2 The distribution of maximum sharp ratio if sampling 10,000 compared with that with 100,000 sampling

take the transaction fees into consideration, it may even support the preference of choosing those portfolios consisting of small number of stocks. The maximum sharp ratio across trading days can vary a lot (Fig. 1B) and its volatility is highly correlated with the market performance (Fig. 1C).

To check how the number of sampling could affect the possibility of capturing the maximum sharp ratio, we run another experiment with 100,000 sampling and the comparison is shown in Fig. 2. With the increased number of sampling, the



Relationship between maximum sharp ratio and number of stocks

Dynamics of maximum sharp ratio across the whole year

Fig. 3 Portfolio analysis using high-frequency data

distribution of maximum sharp ratio does shift right a little. Overall the difference is not significant. Considering the big increase of computational cost when increasing the sampling from 10,000 to 100,000, we kept the sampling to 10,000 in the following experiments.

3.2 Portfolio Optimization with High-Frequency Data

We downloaded the intra-day stock prices from the TAQ database and synchronized the time point. With high-frequency data, we estimated the variance matrix between the 30 stocks for each day. Then we performed the same workflow again. The relationship between the maximum sharp ratio and the number of stocks shown in the low frequency data still holds (Fig. 3a) though in someday the maximum sharp ratio keep increasing or decreasing or stable when changing the number of stocks in the portfolio.

The maximum sharp ratio across trading days (Fig. 3b) also varies a lot since the co-volatilities between stocks are more dynamic. Figure 3b also shows volatility clustering, leverage effects, and the non-stationary process of the stock price movement. The portfolio based on high-frequency data can achieve much higher sharp ratio during the high volatility range of the whole market when compared with Fig. 1b. This relatively strong performance is universal for the global minimum variance portfolio consisting of from 2 stocks to 30 stocks (Fig. 3a). From the comparison of the distribution of those larger than zero maximum sharp ratios between high-frequency data and low frequency data (Fig. 4), we could see that sharp ratios from high-frequency data have a more disperse range and right shift

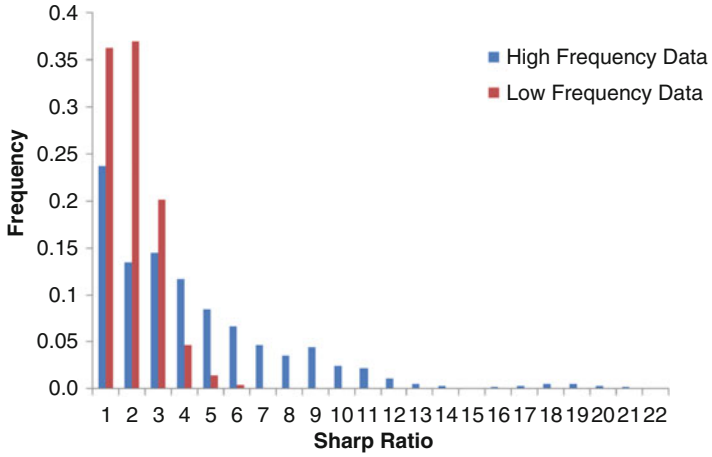


Fig. 4 The distribution of maximal sharp ratios from high- and low-frequency data

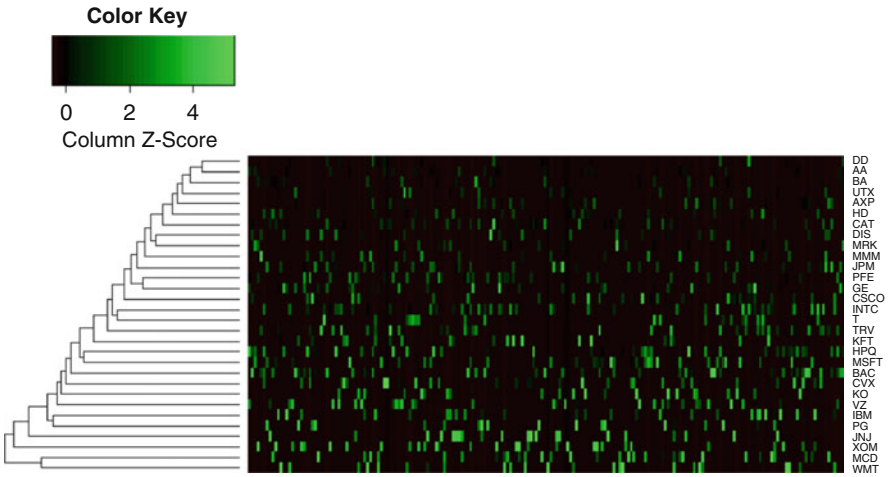


Fig. 5 The composite of the maximum sharp ratio portfolio of five stocks: from left to right is the trading day at 2011 and from top to bottom lists the Dow30 stocks. The value of the heat map is scaled value based on the weight of each stock in the portfolio.

towards higher values than those from low frequency data. It attributes to the fact that high-frequency data could capture the true co-volatilities between stocks.

We also checked the composite of the maximum sharp ratio portfolio. Figure 5 shows an example of the maximum sharp ratio portfolio of five stocks. The composite changes across trading days, indicating rebalance is needed to keep maximum sharp ratio. Nevertheless, some interesting patterns conform to our existing knowledge. For example for the technical stocks, GE, CSCO, INTC, and T closely clustered together. Also PG and JNJ, BA and UTX form a cluster implying their similarity in the underlying sector.

4 Discussion and Conclusions

The portfolio optimization under the Markowitz's framework has been applied widely over the years among practitioners. While appealing in its simplicity, the Markowitz's portfolio fails to capture some important characteristics of the capital markets, especially with wide availability of high-frequency financial data nowadays. This paper proposed an innovative methodology to perform asset allocation using high-frequency data. The empirical comparison between portfolios using low-frequency and high-frequency illustrated the efficiency of this method. For technical details, please refer to Zou and Wu (2012). This methodology may be applied to other applications such as transportation and logistics, where determining allocations in a large pool of available resources is crucial. It could also be used in spatio-temporal applications, where there are situations that we want to determine optimal allocation of different monitoring stations.

Acknowledgements We thank the editor and an anonymous referee for helpful comments that improved the quality of this paper.

References

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *J. Amer. Statist. Assoc.*, 96(453):42–55.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536.

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(2):253–280.

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Econometric analysis of realized covariation: high frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72(3):885–925.

Fan, J., Li, Y., and Yu, K. (2012a). Vast volatility matrix estimation using high frequency data for portfolio selection. *J. Am. Stat. Assoc.* To appear.

Fan, J. and Wang, Y. (2007). Multi-scale jump and volatility analysis for high-frequency financial data. *J. Amer. Statist. Assoc.*, 102(480):1349–1362.

Fan, J., Zhang, J., and Yu, K. (2012b). Asset allocation and risk assessment with gross exposure constraints for vast portfolios. *J. Am. Stat. Assoc.* To Appear.

Hayashi, T. and Yoshida, N. (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 11(2):359–379.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 58:1651–1684.

Mancino, M. E. and Sanfelici, S. (2008). Robustness of Fourier estimator of integrated volatility in the presence of microstructure noise. *Comput. Statist. Data Anal.*, 52(6):2966–2989.

Markowitz, H. (1959). *Portfolio Selection: Ecient Diversication of Investments*. John Wiley & Sons, New York.

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance*, 7:77–91.

Sharpe, W. F. (1966). Mutual fund performance. *Journal of Business*, 39(1):119–138.

Tao, M., Wang, Y., Yao, Q., and Zou, J. (2011). Large volatility matrix inference via combining low-frequency and high-frequency approaches. *J. Amer. Statist. Assoc.*, 106(495):1025–1040.

Wang, Y. and Zou, J. (2010). Vast volatility matrix estimation for high-frequency financial data. *Ann. Statist.*, 38(2):943–978.

Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach. *Bernoulli*, 12(6):1019–1043.

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high-frequency data. *J. Amer. Statist. Assoc.*, 100(472):1394–1411.

Zou, J. and Wu, Y. (2012). Large portfolio allocation using high-frequency financial data. Manuscript.

Applications of Functional Dynamic Factor Models

Spencer Hays, Haipeng Shen, and Jianhua Z. Huang

Abstract Accurate forecasting of zero coupon bond yields for a continuum of maturities is paramount to bond portfolio management and derivative security pricing. Yet a universal model for yield curve forecasting has been elusive, and prior attempts often resulted in a tradeoff between goodness-of-fit and consistency with economic theory. To address this, herein we propose a novel formulation which connects the dynamic factor model (DFM) framework with concepts from functional data analysis: a DFM with functional factor loading *curves*. This results in a model capable of forecasting functional time series. Further, in the yield curve context we show that the model retains economic interpretation. We show that our model performs very well on forecasting actual yield data compared with existing approaches, especially in regard to profit-based assessment for an innovative trading exercise. We further illustrate the viability of our model to applications outside of yield forecasting.

S. Hays (✉)
Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, VA 23284, USA
e-mail: shays@vcu.edu

H. Shen
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: haipeng@email.unc.edu

J.Z. Huang
Department of Statistics, Texas A&M University, College Station, TX 77843, USA
e-mail: jianhua@stat.tamu.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_3,

1 Introduction

The yield curve is an instrument for portfolio management and for pricing synthetic or derivative securities (Diebold and Li 2006). Our contribution to the yield literature is pragmatic: we introduce a dynamic factor model with functional coefficients which reconciles the theory-based desire to model yield data as a curve with the applied need of accurately forecasting that curve over time.

It is of interest to know the yield for all maturities at each point in time (cross-sectional), as well as for a single maturity as it evolves over time (dynamic). Therefore, bond data have cross-sectional and dynamic correlational behaviors to consider for predictive modeling.

Yield curve models have traditionally assumed either of a cross-sectional or time series formulation. The former approach conforms to economic theory; for a given date the emphasis is on fitting a yield curve to existing yields based on no-arbitrage principles. The latter approach is the so-called equilibrium or affine-class models where time series techniques are used to model the dynamics of yield on a short-term or instantaneous maturity. Yields for longer maturities are then derived using an affine model. In any sense, goodness of fit is paramount in a model for it to be of any use. Still, a yield model should be consistent with its underlying theory and maintain a degree of economic interpretation. Cross-sectional/no-arbitrage models ignore the dynamics of yields over time; time series models place emphasis on dynamics at the expense of theory.

Proposed in this chapter is a synthesis of the cross-sectional and dynamic considerations mentioned above. We approach yield curves as a *functional* time series; the yields of the observed maturities are a discrete sampling from a true underlying yield *curve*. To this end we conflate concepts from functional data analysis (Ramsay and Silverman 2002; 2005, FDA; e.g.,) and from dynamic factor analysis/modeling (Basilevsky 1994, DFM; e.g.,). Ours is a dynamic factor model with *functional coefficients* which we call the functional dynamic factor model (FDFM). These functional coefficients, or, *factor loading curves* (FLCs) are natural cubic splines (NCS): a significant result which facilitates interpolation of yields both within and out of sample so that forecasts are indeed true yield curves. While the factor loadings account for the cross-sectional/curve dimension of yields, the dynamic factors, in turn, determine the evolution of these functions over time. Thus, they account for the time series and cross-correlational nature of yield data.

Combining elements from functional data analysis (FDA) and dynamic factor modeling (DFM) to forecast time series of yield curves is not necessarily a new thing. Diebold and Li (2006) developed their Dynamic Nelson–Siegel model (DNS) with functional coefficients which has spawned multiple variants (i.e., Koopman et al. 2010). Also, Bowsher and Meeks (2008) use NCS in conjunction with cointegrated latent factors. Though our FDFM is in a similar vein as those of the aforementioned yield models, there are two notable distinctions. First, the FDFM functional coefficients are estimated; thus, they are free to vary with the particular application to explain the functional nature of the data. Second, estimation of the FDFM is achieved in a single step.

Presented in this chapter is our FDFM which we show to perform very well in regard to yield curve forecasting and also satisfies our two aforementioned criteria: goodness of it and economic interpretability. Further, we do so in multiple assessments which highlight the model's capability of accurately forecasting in both the statistical and financial (profit) sense.

2 Functional Dynamic Factor Models

Consider a general process of a time series of curves $\{x_i(t) : t \in \mathcal{T}; i = 1,\ldots,n\}$, where $\mathcal{T}$ is some continuous interval and i indexes discrete time. We hypothesize that each curve is composed of a forecast-able smooth underlying curve, $y_i(t)$, plus an error component, $\epsilon_i(t)$, i.e.,

$$x_i(t) = y_i(t) + \epsilon_i(t). \tag{2.1}$$

In practice, only a discrete sampling of each curve is observed. Specifically, for $t \in \mathcal{T}$, consider a sample of discrete points $\{t_1, t_2, \ldots, t_m\}$ with $t_j \in \mathcal{T}$ for $j \in \{1,\ldots,m\}$. Then denote $x_{ij} \equiv x_i(t_j)$ as an observed data at time i evaluated at t_j. In terms of forecasting, even if the cross-sectional dimension m is small enough for VAR or DFM methods, there is nothing inherent or implicit in those frameworks to provide direction in terms of forecasting an entire function for all $t \in \mathcal{T}$. Thus, to forecast the smooth curve $y_{n+h}(t)$ for some forecast horizon $h > 0$, we can synthesize the DFM framework with methods from FDA.

2.1 The Model

Via this synthesis, we propose a model referred to as the FDFM. The formulation is similar to that of a DFM where the observed data $\{x_{ij}\}$ is a function of a small set of K latent dynamic factors $\{\beta_{ik}; k = 1,\ldots,K\}$ and their corresponding factor loadings. But in this setting the factor loadings $f_{kj} \equiv f_k(t_j)$ are discrete samples from continuous, unobserved though nonrandom FLCs $f_k(\cdot)$. Together, the dynamic factors with their functional coefficients generate the forecast-able part of the time series of curves $\{x_i(t)\}$.

In theory, the dynamic factors can follow any type of time series process such as (V)ARIMA, but for the purpose of this method we focus on factors with stationary and independent AR(p) errors. These factors can include explanatory variables or just a constant; here we will consider only the latter. A complete discussion of the former can be found in Hays et al. (2012a;b); Judge (1985) provides a good reference for the procedure in general.

The order of the AR processes may differ among factors, but for notational simplicity here we assume it is the same for all. For the model to be identified,

we require that the functional coefficients are orthonormal, although other types of constraints may be employed to ensure identification (i.e., conditions on the covariance functions). The model is explicitly stated as

$$\begin{cases} x_i(t_j) = \sum_{k=1}^{K} \beta_{ik} f_k(t_j) + \epsilon_i(t_j), \\ \beta_{ik} = \sum_{r=1}^{p} \varphi_{rk} \beta_{i-r,k} + v_{ik}, \\ \int_T f_k(t) f_l(t) dt = \begin{cases} 1 \text{ if } k = l, \\ 0 \text{ otherwise,} \end{cases} \end{cases} \tag{2.2}$$

with $\epsilon_i(t_j) \equiv \epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $v_{ik} \overset{i.i.d.}{\sim} N(0, \sigma_k^2)$ and $E[v_{ik}\epsilon_{i'j}] = 0$ for $i, i' = 1, \ldots, n$. This is a broad framework that includes the standard versions of both DFMs and FPCA models: when the coefficients $\{f_k(t)\}$ are nonfunctional, Model (2.2) reduces to the standard DFM; when the factors $\{\beta_k\}$ are non dynamic, the model is similar to FPCA.

2.2 Estimation

With the error assumptions for Model (2.2) we propose estimation via maximum likelihood (ML). To ensure smooth and functional estimates for the FLCs, we augment the likelihood expression with "roughness" penalties (Green and Silverman 1994) and maximize a *penalized* likelihood expression. Because our dynamic factors are unobserved, we consider this a problem of missing data and use the expectation maximization (EM) algorithm (Dempster et al. 1977) to estimate model parameters and smooth curves.

2.2.1 Penalized Likelihood

Let the $n \times m$ matrix $\mathbf{X}$ denote collectively the observed data where the (i, j)th element of $\mathbf{X}$ is x_{ij} for $i = 1, \ldots, n$, $j = 1, \ldots, m$. The rows of $\mathbf{X}$ correspond to yield curves for a fixed date; the columns are the time series of yield for a specific maturity. Next, we denote $f_{kj} = f_k(t_j)$, the *row* vector $\mathbf{f}_k = [f_{k1}, \ldots, f_{km}]$, and $\mathbf{F}' = [\mathbf{f}_1', \ldots, \mathbf{f}_K']$. In a similar manner, we define $\beta_k = [\beta_{1k} \ldots \beta_{nk}]'$ and the matrix $\mathbf{B}_{n \times K} = [\beta_1 \ldots \beta_K]$. Thus the columns of $\mathbf{B}$ are the time series factors $\beta_1, \ldots, \beta_K$. Then, the Model (2.2) is represented in matrix form as

$$\mathbf{X}_{n \times m} = \mathbf{B}_{n \times K} \mathbf{F}_{K \times m} + \epsilon_{n \times m} = \sum_{k=1}^{K} \beta_k \mathbf{f}_k + \epsilon, \tag{2.3}$$

where $\epsilon = [\epsilon_{ij}]_{n \times m}$ with $\epsilon_{ij} = \epsilon_i(t_j)$.

To derive the log-likelihood expression, we rely on successive conditioning of the joint distribution for $\mathbf{X}$ and $\mathbf{B}$:

$$l(\mathbf{X},\mathbf{B}) = l(\mathbf{B}) + l(\mathbf{X}|\mathbf{B}). \tag{2.4}$$

We have assumed that the K factors of AR(p) series are independent, thus their joint distribution is the product of the individual distributions. To each of those, we further condition on the first p values of each factor time series; thus our likelihood (2.4) is a *conditional* one.

To ensure each $f_k(\cdot)$ is smooth, following Green and Silverman (1994) we include roughness penalties in solving for the K FLCs $\mathbf{f}_k$. These terms place a condition on the second derivative of each function $f_k(\cdot)$ to ensure that the function is not too "rough." Combining the likelihood expression with the roughness penalties, we obtain the following penalized log-likelihood:

$$l_p(\mathbf{X},\mathbf{B}) = l(\mathbf{B}) + l_p(\mathbf{X}|\mathbf{B}) \equiv l(\mathbf{B}) + \left[l(\mathbf{X}|\mathbf{B}) + \sum_{k=1}^{K} \lambda_k \int \left[f_k''(t)\right]^2 dt \right]. \tag{2.5}$$

The penalty parameter λ_k controls how strictly the roughness penalty is enforced, and we allow it to differ for each loading curve (discussed in Sect. 2.4). We refer to the latter term in (2.5), $l_p(\mathbf{X}|\mathbf{B})$, as the penalized sum of squares (PSS). Intuitively, optimization of PSS balances a familiar goodness-of-fit criterion with a smoothness requirement for the resulting estimates of $f_k(t)$.

Thus, to estimate the model, we will optimize the penalized, conditional log-likelihood $l_p(\mathbf{X},\mathbf{B})$ with respect to the set of parameters and FLCs. Because $\mathbf{B}$ are unobserved, we treat their absence as a missing value problem and use the expectation maximization algorithm (Dempster et al. 1977) to optimize Expression (2.5). Before this is discussed in Sect. 2.2.3, we will currently assume the dynamic factors are known.

To estimate the loading curves $f_k(t)$, according to Theorem 2.1 of Green and Silverman (1994), for fixed k, the minimizer $\hat{f}_k(\cdot)$ of PSS is a natural cubic spline with knot locations $t_1,\ldots,t_m$. Further, this NCS interpolates the discrete row vector $\hat{\mathbf{f}}_k$ which is the solution to the minimization problem

$$\min_{\mathbf{f}_k} \left[l(\mathbf{X}|\mathbf{B}) + \lambda_k \mathbf{f}_k \Omega \mathbf{f}_k' \right], \tag{2.6}$$

where $\Omega_{m\times m}$ is a matrix determined solely by the spline knot locations.

2.2.2 Connection with Natural Cubic Splines

The origin of the penalty matrix Ω from (2.6) stems from Green and Silverman (1994). Conceptually, functions of first and second differences of each f_{kj} serve as derivatives of $f_k(t)$ evaluated at t_j; see Hays et al. (2012a) for details. The important

concept is that the matrix Ω depends only on the observed values for t_j. This facilitates a rather useful result based on Theorem 2.1 of Green and Silverman (1994).

Proposition 2.1. *For fixed k, the* $\hat{f}_k(\cdot)$ *optimizing* PSS *in* (2.5) *is a natural cubic spline with knot locations at* t_j*, and* $\int \left[f_k''(t)\right]^2 dt = \mathbf{f}_k \Omega \mathbf{f}_k'$.

2.2.3 EM Algorithm

First introduced by Dempster et al. (1977), then refined by Meng and Rubin (1993), the EM is an iterative method by which to impute missing data with values based on conditional expectations using the observed data. First, the EM is inaugurated with initial values for the factors and FLCs. From these initial values, maximum likelihood estimates for the parameters are calculated based on (2.5); we call this *Step 0*. Then the algorithm alternates between the E-step and the M-step. In the *E-step*, values for the factor time series are calculated as conditional expectations given the observed data and current values for the MLEs. In the *M-step*, MLEs are calculated for the FLCs and other parameters based on the factor scores from the conditional expectations in the E-step. After the initial step, the E-step and the M-step are repeated until differences in the estimates from one iteration to the next are sufficiently small.

Step 0: Akin to the method used in Shen (2009), initial values for $\mathbf{B}$ are composed of the first K singular values and left singular vectors from the singular value decomposition (SVD) of the data matrix $\mathbf{X}$. Initial values for $\mathbf{F}$ are the corresponding right singular vectors. From these, initial parameter estimates are computed for σ^2 and the set of factor parameters $\{\sigma_k^2, c_k, \varphi_{1,k}, \ldots, \varphi_{p,k}\}$.

The E-Step: Derivation of the conditional moments for the E-Step requires the expressions of some of the unconditional moments. Let $\mathbf{X} \equiv vec(\mathbf{X})$ and $\beta \equiv vec(\mathbf{B})$. Then (2.3) can be rewritten as

$$X = (\mathbf{F}' \otimes \mathbf{I}_n)\beta + vec(\epsilon) = \sum_{k=1}^{K} (\mathbf{f}_k \otimes \mathbf{I}_n)\beta_k + vec(\epsilon). \tag{2.7}$$

Define the $n \times n$ variance matrix for β_k as Σ_k, and let $\mathbf{c}$ be the $K \times 1$ vector with elements $c_k/[1 - (\sum_{r=1}^{p} \varphi_{r,k})]$. Then

$$\begin{aligned} E[\beta] \equiv \mu_\beta = \mathbf{c} \otimes \mathbf{1}_n \; E[X] \equiv \mu_{\mathbf{X}} = (\mathbf{F}' \otimes \mathbf{I}_n)\mu_\beta \\ Var[\beta] \equiv \Sigma_\beta = diag\{\Sigma_1, \ldots, \Sigma_K\} \; Cov[\beta, X] \equiv \Sigma_{\beta,\mathbf{X}} = \Sigma_\beta(\mathbf{F} \otimes \mathbf{I}_n) \\ Var[X] \equiv \Sigma_{\mathbf{X}} = (\mathbf{F}' \otimes \mathbf{I}_n)\Sigma_\beta(\mathbf{F} \otimes \mathbf{I}_n) + \sigma^2 \mathbf{I}_{nm}. \end{aligned} \tag{2.8}$$

Next, using properties of multivariate normal random vectors, the conditional distribution of $\beta|\mathbf{X}$ can be found. Let

$$\begin{pmatrix}\beta\\ X\end{pmatrix} \sim N\left[\begin{pmatrix}\mu_\beta\\ \mu_\mathbf{X}\end{pmatrix}, \begin{pmatrix}\Sigma_\beta & \Sigma_{\beta,\mathbf{X}}\\ \Sigma_{\mathbf{X},\beta} & \Sigma_\mathbf{X}\end{pmatrix}\right].$$

Then

$$\begin{cases}\mu_{\beta|\mathbf{X}} \equiv E[\beta|X] = \mu_\beta + \Sigma_{\beta,\mathbf{X}}\Sigma_\mathbf{X}^{-1}(X-\mu_\mathbf{X}),\\ \Sigma_{\beta|\mathbf{X}} \equiv Var[\beta|X] = \Sigma_\beta - \Sigma_{\beta,\mathbf{X}}\Sigma_\mathbf{X}^{-1}\Sigma_{\mathbf{X},\beta},\\ E[\beta\beta'|\mathbf{X}] = \Sigma_{\beta|\mathbf{X}} + \mu_{\beta|\mathbf{X}}\mu'_{\beta|\mathbf{X}}.\end{cases} \tag{2.9}$$

From a computational standpoint there is concern over the inversion of $\Sigma_\mathbf{X}$ which is of order nm. However it can be shown (Hays et al. 2012a, i.e., see) that a Sherman–Morrison–Woodbury factorization (Press et al. 1992, e.g.,) reduces to K, $n \times n$ inversions rather than a single from an $nm \times nm$ inversion.

With the conditional moments, the E-step of the EM posits that the latent time series factors are replaced with the known values of the conditional distribution given $\boldsymbol{X}$. Values for terms like β_{ik} come directly from the vector $\mu_{\beta|\mathbf{X}}$ and cross-factor terms like $\beta_{ik'}\beta_{hk}$ are obtained from the $E[\beta\beta'|X]$ matrix.

The M-Step: For each EM iteration, the M-step optimizes the conditional penalized log-likelihood in (2.5) given the observed data and the current parameter estimates for Θ. The M-step, then, is just a matter of making these substitutions into the likelihood and solving for the MLEs. After the M-Step, we return to the E-Step to update the values for the factor time series. This procedure is repeated until the parameter estimates from one iteration of the EM are sufficiently close to those of the next.

2.3 *Forecasting and Curve Synthesis*

Forecasting is straightforward: for illustrative purposes, suppose we estimate our FDFM with K factors following an AR(1) process with constants $\{c_k\}$, $k = 1,\ldots,K$. Then the h-step ahead forecasted curve $\hat{x}_{n+h|n}(t)$ is based on the components of the forecast of the factor time series $\hat{\beta}_{n+h|n,k}$ and the estimated FLCs $\hat{f}_k(t)$:

$$\begin{cases}\hat{x}_{n+h|n}(t) = \sum_{k=1}^K \hat{\beta}_{n+h|n,k}\hat{f}_k(t)\\ \hat{\beta}_{n+h|n,k} = \hat{c}_k + \hat{\varphi}_k\hat{\beta}_{n+h-1,k} = \sum_{r=0}^{h-1}\hat{\varphi}^r\hat{c}_k + \hat{\varphi}_k^h\beta_{nk}.\end{cases} \tag{2.10}$$

The NCS result of Sect. 2.2.2 ensures we can interpolate $\hat{f}_k(t)$ to any degree of fineness between any two knot locations t_j and t_{j+1}. See (Green and Silverman

1994) for the specific formulation. It is worth noting that for $t < t_1$, or $t > t_m$, the $\hat{f}_k(t)$ is a linear extrapolation. Using this method together with (2.10) we can impute *and* forecast at the same time; a result that enables, e.g., yield forecasts for bonds of maturities that *have not been observed*.

2.4 Computational Efficiency

The estimation algorithm using the iterative EM with this FDFM formulation can be computationally intensive. However, in addition to the factorization of $\Sigma_{\mathbf{X}}$, two further results hasten computation and convergence. Generalized cross validation (GCV) based on column-deletion of $\mathbf{X}$ can be used to solve for each optimal λ_k and with some algebraic manipulation, this often time-consuming iterative method reduces to a rather time efficient algorithm.

In the M-step, when products of the factors appear, such as $\langle \beta_k, \beta_h \rangle = E[\langle \beta_k, \beta_h \rangle | X]$, then the imputation comes from the $E[\beta\beta' | \mathbf{X}]$ matrix. It can be shown that $\Sigma_{\beta|\mathbf{X}}$ is block diagonal; this property facilitates a rather convenient result regarding between-factor cross products. It can be shown that the conditional expectation of a product of two (distinct) factors is simply the product of their individual expectations. This greatly simplifies the M-step calculations. See Hays et al. (2012a;b) for details.

3 Application to Yield Curve Data

In this section we consider the application of our FDFM to actual yield data. We use the same data as Diebold and Li (2006)[1] which is a sample of monthly yields on zero coupon bonds of 18 different maturities (in months):

$$t_j = \{1.5, 3, 6, 9, 12, 15, 18, 21, 24, 30, 36, 48, 60, 72, 84, 96, 108, 120\},$$

$j = 1, \ldots, m = 18$. The yields are from the period January 1985 through December 2000 (192 months) originally obtained from forward rates provided by the Center for Research in Securities Prices (CRSP), then converted to unsmoothed Fama-Bliss yield rates; see Fama and Bliss (1987) for further details on the method.

[1] See http://www.ssc.upenn.edu/~fdiebold/papers/paper49/FBFITTED.txt for the actual data.

3.1 Candidate Models

In the following sections we compare the FDFM with the DNS model presented in Diebold and Li (2006). Their model is composed of three factors with corresponding FLCs. The factor loadings are pre-specified parametric curves (see the dashed curves in Fig. 1) based on financial economic theory. Let $x_i(t)$ denote the yield at date i on a zero coupon bond of maturity t, then the DNS model is represented as

$$\begin{cases} x_i(t) = \sum_{k=1}^{3} \beta_{i,k} f_k(t) + \epsilon_i(t), \text{ for } i = 1,\dots,n, \\ f_1(t) \equiv 1, f_2(t) \equiv \frac{1-\exp(-\alpha_i t)}{\alpha_i t}, f_3(t) \equiv f_2(2) - \exp(-\alpha_i t), \\ \beta_{i,k} = c_k + \varphi_k \beta_{i-1,k} + \zeta_{i,k}, \text{ for } k = 1,2,3, \end{cases} \tag{3.1}$$

evaluated at maturities t_j, $j = 1,\dots,m$. The first loading curve $f_1(t)$ is constant and intended to represent the long-term component of yields (level); the second $f_2(t)$ represents a short-term component, or slope. Finally, the third loading $f_3(t)$ represents a mid-term component, or curvature. The parameter α_i determines the point $t^*(\alpha_i)$ at which $f_3(t)$ achieves its maximum. While this can be estimated as a fourth factor (see, e.g., Koopman et al. 2010), Diebold and Li (2006) set α_i to a fixed value for all $i = 1,\dots,n$. This results in entirely predetermined, parametric curves. The specific value $\alpha = 0.0609$ is determined by their definition of "mid-term" as $t = 30$ months.

Estimation of the DNS model is a two-step procedure. First, time series of factor scores of $\hat{\beta}_{i,k}$ are estimated by ordinary least squares (OLS) of $x_i(t_j)$ on $[1, f_2(t_j), f_3(t_j)]$ for $j = 1,\dots,m$ at each time point $i = 1,\dots,n$. Second, an AR(1) model is fit on each series $\hat{\beta}_{i,k}$ for the purpose of forecasting $\hat{\beta}_{n+1,k}$ and ultimately $\hat{x}_{n+1}(t_j)$ via (2.10) from Sect. 2.3.

3.2 Assessment

We assess the performance of the FDFM in three distinct exercises. The first two are traditional error-based assessments of forecasts or within-sample predictions of yield curves or sections thereof. The final application is a combination of both forecasting and curve synthesis. Through an adaptation of the trading algorithms introduced in Bowsher and Meeks (2008), we develop trading strategies based on the forecasts of the FDFM and DNS models and assess the resulting profit generated by each.

For each of these, as a comparison, we use the DNS specification aforementioned above in Sect. 3.1. For the purpose of making an unbiased comparison, we use a similar formulation of our FDFM model with 3 factors following independent AR(1) processes. The key distinction between this FDFM model and the DNS model is that the FDFM estimates the model simultaneously, while the estimation for the DNS model requires two steps.

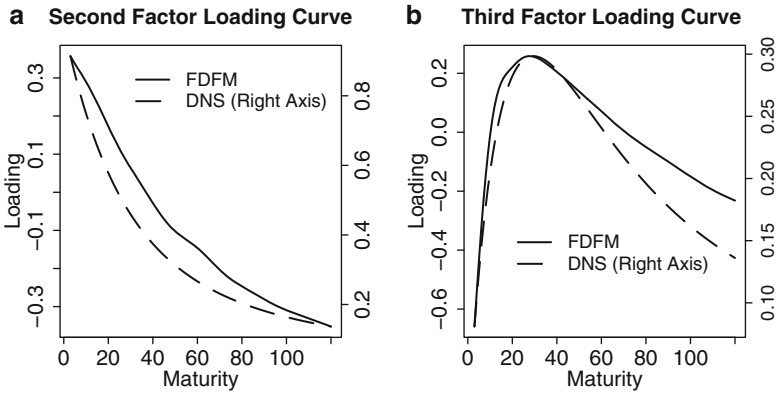


Fig. 1 Example of factor loading curves: FDFM curves (solid, left axis) estimated from the period May 1985 to April 1994; pre-specified DNS curves (dashed, right axis). FDFM estimates closely resemble the shape of the DNS curves for the second and third factors. Dual axes have been utilized to account for difference in scale.

Figure 1, Panels (a) and (b), show an example of the FLCs estimated by the FDFM (solid line) for the period May 1985 through April 1994. Pictured alongside, the dashed line plots the DNS model curves. Recall the DNS motivation for the form of f_1, f_2, and f_3 was an economic argument, while the formulation of the FDFM described in Sect. 2 is based entirely on statistical considerations. Despite this, we see that the FDFM model is flexible enough to adapt to a specific application. FLCs $f_2(t)$ and $f_3(t)$ from the FDFM assume the behavior of those from the DNS model without imposing any constraints that would force this. Thus, the FDFM inherits the economic interpretation of $f_2(t)$ and $f_3(t)$ set forth in Diebold and Li (2006). We consider the first FDFM factor loading curve (not pictured) as the mean yield.

3.2.1 Forecast Error Assessment

In this section, we compare the FDFM and DNS models using a rolling window of 108 months to forecast the yield curve 1, 6, or 12 months ahead. Yields on bonds of maturity less than 3 months are omitted in order to match the methodology used in Diebold and Li (2006). To compare the models we use the root mean squared forecast error (RMSFE):

$$\text{RMSFE}_j = \sqrt{\sum_{i=1}^{r} \frac{\left[x_{n+h}(t_j) - \hat{x}_{n+h}(t_j)\right]^2}{r}}.$$

where $r = 84, 79, 73$ is the number of rolling forecasts for forecast horizon $h = 1, 6, 12$, respectively.

Table 1 RMSFE: 1, 6, and 12 month ahead Yield Curve Forecast Results. The better result between the two models is highlighted in **bold**. RMSFE is typically lower with the FDFM for 1, 6 and 12 months ahead.

	RMSFE					
	1 Month Ahead		6 Months		12 Months	
Maturity	DNS	FDFM	DNS	FDFM	DNS	FDFM
3 Months	0.176	**0.164**	**0.526**	0.535	0.897	**0.867**
1 Year	0.236	**0.233**	**0.703**	0.727	0.998	**0.967**
3 Years	0.279	**0.274**	0.784	**0.775**	1.041	**0.947**
5 Years	0.292	**0.277**	0.799	**0.772**	1.078	**0.953**
10 Years	0.260	**0.250**	0.714	**0.697**	1.018	**0.921**

Table 2 Average RMSFE; FDFM as a fraction of DNS

Omitted	1	2	3	4	5	6	7	8
Short	**0.88**	**0.95**	**0.94**	**0.87**	**0.99**	**0.99**	1.34	**0.92**
Mid	**0.97**	**0.90**	**0.98**	**0.93**	1.01	**1.00**	1.11	1.17
Long	1.05	1.13	1.06	1.64	**0.88**	1.67	**0.93**	1.82
All	**0.95**	**1.00**	**0.98**	1.14	**0.95**	1.26	1.13	1.39

A summary of the forecasting performance is shown in Table 1. For month ahead forecasts, the RMSFE is lower for all five displayed maturities. For 6 months ahead, DNS outperforms FDFM just two out of five times. For 12-month ahead forecasts, FDFM has lower RMSE for all five maturities.

3.2.2 Curve Synthesis

Because each factor loading curve $\hat{f}_k(\cdot)$ is an NCS, between any two observed maturities t_j and t_{j+1}, we can calculate the value for $\hat{f}_k(t)$. It follows, then, that between any two time series of yields $\{x_i(t_j)\}_{i=1}^n$ and $\{x_i(t_{j+1})\}_{i=1}^n$, we are able to replicate an entire time series for the intermediate maturity t: $\{\hat{x}_i(t)\}_{i=1}^n$.

To illustrate this point, we use the entire data set (see introduction of Sect. 3), i.e., use $i = 1,\ldots,n = 192$ months of yield data for maturities t_j, $m = 18$. For both the DNS and FDFM models, we delete a set of $l = 1,\ldots,8$ adjacent time series from the data, estimate the model, then assess the prediction RMSFE of the predicted series in reference to the actual deleted series. More information on the RMSFE calculation can be found in the original paper Hays et al. (2012a).

Results are presented in Table 2 with FDFM as a fraction of DNS. In general, as L increases we see the expected decline in the performance of the FDFM model relative to DNS. The average RMSFE on short-term bonds for the FDFM remains surprisingly robust as we delete more and more maturities. On mid-term bonds, DNS results in lower prediction error when the number of deleted series reaches 5 or more. For long term, DNS more or less outperforms FDFM across the board (this trend will be echoed in Sect. 3.2.3).

3.2.3 Portfolio-Based Assessment

RMSFE-type assessment provides a good diagnostic measure of forecast performance from a statistical perspective. In this section we consider an adaptation of the profit-based assessment introduced in Bowsher and Meeks (2008). There they considered three separate trading strategies and calculated cumulative profits over the trading period. Here we present results for one of their strategies and comment on the other two. Complete results for those can be found in the main paper of Hays et al. (2012a). In any case, evaluation is based on the same concept: accurate prediction of the expected spread between long and short positions. This serves as a good capstone exercise for our presentation of the FDFM as it simultaneously involves both forecasting *and* curve synthesis: the primary uses for our model.

We use the same rolling window of 108 months as described in Sect. 3.2.1 so that the trading algorithm is employed every month over the course of 84 months. Each period i we create a portfolio consisting of a \$1M purchase of one bond or set of bonds and a corresponding sale of another bond or set of bonds for the same amount. Therefore, the net investment per period is \$0. The decision of which bond to sell and which to buy is made based on the sign of the predicted spread in their one period returns.

At time $i+1$ we cash out our portfolio and record the cumulative profit over the 84-month trading period. Denoting the yield at time i of a zero coupon bond of maturity t months as $x_i(t)$, the price of the bond at time i is

$$P_i(t) = \exp[-tx_i(t)]. \tag{3.2}$$

Correspondingly, the price the next period (month) is then $P_{i+1}(t-1) = \exp[-(t-1)x_{i+1}(t-1)]$ since in the month that has elapsed the maturity is reduced by 1 month. We denote the one period return as

$$R_{i+1}(t) = \left[\frac{P_{i+1}(t-1)}{P_i(t)}\right] - 1, \tag{3.3}$$

and the log one period return as $r_{i+1}(t) \equiv \ln[1+R_{i+1}(t)]$. Equations (3.2) and (3.3) imply

$$r_{i+1}(t) = tx_i(t) - (t-1)x_{i+1}(t-1). \tag{3.4}$$

Thus for a forecasted yield $\hat{x}_{i+1|i}(t)$ we have $\hat{r}_{i+1|i}(t) = tx_i(t) - (t-1)\hat{x}_{i+1|i}(t-1)$, which is a combination of both actual and forecasted yields. We use the data presented in the beginning of Sect. 3 and thus are limited to a set of non-consecutive observed maturities. Akin to Bowsher and Meeks (2008), we rely on linear interpolation of $x_i(t-1)$ to provide the yield for $x_i(t)$ and use the same random walk forecast (RW) as a benchmark by which to compare models:

$$x_{i+1}(t) = x_i(t) + \eta_{i+1}(t),\ \eta_{i+1}(t) \overset{iid}{\sim} WN(0, v^2), \tag{3.5}$$

with forecast $\hat{x}_{i+1|i}(t) = x_i(t)$.

Table 3 Optimal Pairs Portfolio.

		Profit ($000)					Profit ($000)		
	t_1	FDFM	DNS	RW		t_1	FDFM	DNS	RW
Short	3	1,013	**3,574**	−228	Mid	21	**1,246**	202	680
	6	1,381	**2,828**	−133		24	**1,592**	242	70
	9	**1,061**	1,013	−297		30	**2,284**	203	−951
	12	**1,873**	−367	−307		36	**1,466**	919	−173
	15	**1,519**	−582	−432	Long	48	−361	**589**	236
	18	**1,081**	−481	−263		60	**740**	339	−284
						72	−131	−1	**72**

Optimal Pairs Algorithm. Our main strategy of focus is the most sophisticated method by creating portfolios of an optimal pair of bonds each period i. Given a fixed value of t_1, we choose t_{2i} to optimize the absolute spread in predicted return:

$$t_{2i} = \arg\max_{t \neq t_1} |\hat{r}_{i+1|i}(t) - \hat{r}_{i+1|i}(t_1)|. \tag{3.6}$$

Because we examine multiple portfolios, we use a sparser set of maturities in this exercise than previous, though of the same range. This set is defined by the observed maturities in the beginning of Sect. 3: $t_1, t_{2i} \in T = \{4, 7, 10, 13, 16, 19, 22, 25, 31, 37, 49, 61, 73, 85\}$. We perform this exercise for all choices of t_1 as long as $t_1 < t_{2i}$ and compare the results. Our investment rule d_i at time i and resulting profit π_{i+1} the next period is:

$$\begin{aligned} d_i &= \$1M \times \text{sgn}[\hat{r}_{i+1|i}(t_{2i}) - \hat{r}_{i+1|i}(t_1)], \\ \pi_{i+1} &= d_i[R_{i+1}(t_{2i}) - R_{i+1}(t_1)] \approx d_i[r_{i+1}(t_{2i}) - r_{i+1}(t_1)]. \end{aligned}$$

We set $d_i = 0$ whenever $\hat{r}_{i+1|i}(t_{2i}) = \hat{r}_{i+1|i}(t_1)$.

The results of the strategy are shown in Table 3. When the choice of t_1 is 6 months or less, the DNS model generates greater cumulative profit than either of the other models. However, when the choice of t_1 is within 9 and 36 months, the FDFM consistently generates significantly greater profit than the DNS and RW models. Thus, when we are free to pick the bond that optimizes the predicted spread each period, the FDFM performs rather well, provided the maturity of the first bond is within a certain range.

Other Algorithms. Every period i we form a portfolio of sub-portfolios with two bonds $\{t_1, t_{2,j}\}$. Define weights w_j as the proportion of the historical absolute excess return on portfolio $\{t_1, t_{2,j}\}$ to the sum over all j of the same. Again cumulative profit is calculated over the same trading period. Use of the FDFM model results in nearly twice the cumulative profit produced from the DNS model. In the interest of space the complete results cannot be displayed here but appear in Hays et al. (2012a).

In the optimal pairs algorithm above, the choice of the optimal second bond can vary from one period to the next so it is not clear what a *consistently*

good combination is. Thus, for the third strategy we consider an exploratory and exhaustive approach as a diagnostic assessment of which combination of bonds our model excels. In this modification of strategy 1 from Bowsher and Meeks (2008), the portfolio is a simple one consisting of two bonds with maturities t_1 and t_2. For the duration of the strategy, these maturities remain fixed over all periods. We examine the cumulative profit of all combinations of this type or portfolio such that $t_2 > t_1$.

Once more, complete results can be found in Hays et al. (2012a), but briefly the FDFM model typically has the greatest profit when $t_2 \in \{30, \ldots, 72\}$. These results are consistent with Sects. 3.2.1 and 3.2.2: the FDFM was either comparable or better on RMSFE for forecasting and for imputation on maturities in this range. We also see a certain similarity in these results to those of the Optimal Pairs Algorithm. Namely, that the FDFM typically outperformed the other two models when t_1 was exactly in this range.

4 Conclusion and Discussion

In this chapter we developed a method for modeling and forecasting functional time series, specifically in the application to yield curve forecasting. This novel approach synthesizes concepts from functional data analysis and DFM culminating in an FDFM. The FDFM is both consistent economic-theory *and* exhibits accurate goodness-of-fit. By specifying error assumptions and smoothness conditions for functional coefficients, estimation by the Expectation Maximization algorithm results in nonparametric FLCs that are NCS'. Thus for a given time series of curves we can forecast entire curves as opposed to a discrete multivariate time series. Indeed, this exciting new class of models is fertile for further development and application.

References

Basilevsky, A. (1994), *Statistical Factor Analysis and Related Methods*, Wiley: New York.

Bowsher, C. G. and Meeks, R. (2008), "The Dynamics of Economic Functions: Modeling and Forecasting the Yield Curve," *Journal of the American Statistical Association*, 103, 1419–37.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via EM Algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, 39, 1–38.

Diebold, F. X. and Li, C. (2006), "Forecasting the Term Structure of Government Bond Yields," *Journal of Econometrics*, 130, 337–64.

Fama, E. and Bliss, R. (1987), "The Information in Long-Maturity Forward Rates," *American Economic Review*, 77, 680–92.

Green, P. J. and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall: New York.

Hays, S., Shen, H., and Huang, J. Z. (2012a), "Functional dynamic factor models with application to yield curve forecasting," *Annals of Applied Statistics*, 6, 870–894.

— (2012b), "Supplement to "Functional Dynamic Factor Models with Application to Yield Curve Forecasting"," *Available online at http://www.unc.edu/£\sim£haipeng/publication/FDFMsupplement.pdf*.

Judge, G. G. (1985), *The Theory and Practice of Econometrics*, Wiley: New York.

Koopman, S. J., Mallee, M. I. P., and der Wel, M. V. (2010), "Analyzing the Term Structure of Interest Rates Using the Dynamic Nelson-Siegel Model with Time-Varying Parameters," *Journal of Business and Economic Statistics*, 28, 329–343.

Meng, X. L. and Rubin, D. B. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992), *Numerical Recipes in Fortran: The Art of Scientific Computing*, Cambridge University Press: New York.

Ramsay, J. O. and Silverman, B. W. (2002), *Applied Functional Data Analysis: Methods and Case Studies*, Springer-Verlag: New York.

— (2005), *Functional Data Analysis*, Springer-Verlag: New York, 2nd ed.

Shen, H. (2009), "On Modeling and Forecasting Time Series of Curves," *Technometrics*, 51, 227–38.

Mixed Modeling for Physician-Direct Campaigns

Wei Huang and Lynda S. Gordon

Abstract Today, pharmaceutical companies leverage multiple channels, such as sales rep visits, samples, professional journals, and even consumer mass media, to reach out to physicians in order to increase product awareness and drug knowledge to gain incremental market share or prescribing volume. Measuring the influence of each individual channel is critical for future planning and resource optimization. The analytic challenge occurs when physicians are exposed to multiple channels simultaneously and the impact of each channel may have different life spans. Traditional ANCOVA (analysis of variance with covariates) is no longer sufficient to see the whole picture. In this paper, we will present a mixed modeling approach to longitudinal data to answer two important business questions: (1) How effective are different channels in promoting sales? (2) How should we allocate resources across multiple channels?

1 Introduction

Pharmaceutical companies rely on product promotions to inform physicians and other health care professionals (HCPs) of the existence and characteristics of new product and to reinforce awareness of established products, especially those facing branded competitors (Duetsch LL 1998). In 2008, pharmaceutical companies spent at least $20.5 billion on promotional activities, excluding expense of the free samples distributed to physicians, or around 10.8% of the US sales reported by the Pharmaceutical Research and Manufacturers of America (PhRMA) (The Congressional Budget Office 2009). In 2011, a projected $29 billion was spent on promotion (SK&A 2012).

W. Huang (✉) • L.S. Gordon
Merkle Inc., 7001 Columbia Gateway Drive, Columbia, MD, 21046 USA
e-mail: whuang@merkleinc.com; lsgordon@merkleinc.com

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_4,

During the past two decades, pharmaceutical companies have significantly expanded their mix of ways to promote their products, first by shifting from heavily relying on sale representative detailing to increased use of multichannel promotions including both personal and nonpersonal and second, by expanding promotional efforts beyond physicians and other HCPs to direct-to-consumer communications consumers. The first change is partially due to harder access to physicians and other health care providers (Grogan 2010) and the second change happened after the Food and Drug Administration (FDA) issued draft regulatory guidance in 1997 allowing pharmaceutical companies to list only major side effects and a toll free number or internet address for details in their television ads (Duetsch LL 1998).

With so many channels, pharmaceutical company marketers need to answer two important questions to optimize their promotion efforts: (1) How effective are different channels in promoting sales? (2) How should we allocate resources across multiple channels?

2 Data Collection

While pharmaceutical companies leverage different channels to reach out to physicians or other health care professionals or consumers, data can be collected from multiple points of interaction between patients, HCPs, and pharmaceutical companies (Fig. 1).

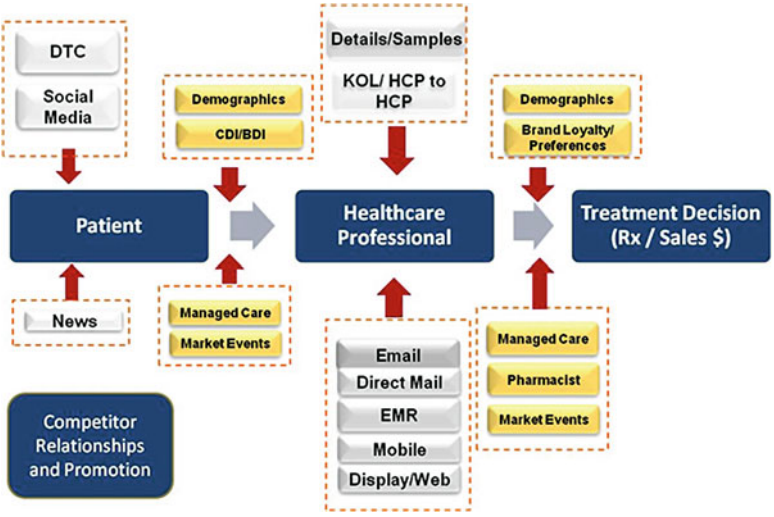


Fig. 1 Communications and interactions: a measurement framework for pharmaceutical brand promotion response and optimization

Companies such as IMS Health or Wolters Kluwer (now Source Healthcare Analytics) provide projected Rx Prescription sales at HCP level for most retail brands. Sales of non-retail, or infusion or administered drugs cannot be tracked by HCP and are not the focus of this paper.

For channels targeting HCPs directly, such as sales representative detailing and sample distribution, data can be collected from internal sales and marketing teams at HCP level, although competitor promotional efforts are not widely available. For channels targeting consumers directly, we convert consumer impressions or reach from these channels to HCP level metrics. For example, we use Gross Rating Point (GRP) within DMA to capture the volume and frequency of TV Ads and use it as model covariates in our analysis.

Physicians' preference may vary significantly. Thus physicians' demographics, such as age and gender, as well as the patient demographics collected at zip-code level based on proximity to HCP offices, could also shed light on behavior and thus need to be included in the analysis.

3 Exploratory Data Analysis

Exploratory data analysis is a critical step in analysis as it helps us: (1) validate the accuracy of data collected and (2) identify potential trend and outliers that can guide model selection. In the example below (Fig. 2), we observe a sharp drop in

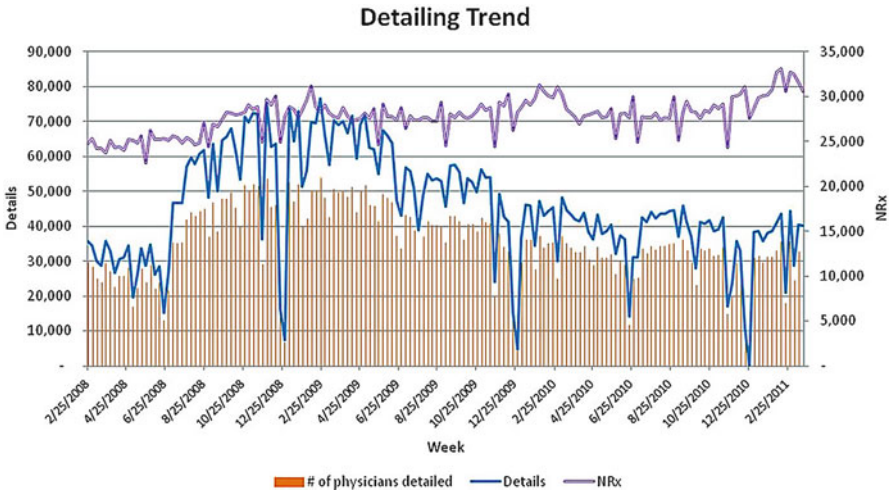


Fig. 2 Example of exploratory data analysis

key weeks every year—this is resulting from sales representatives and HCPs taking days off for holidays and need to be controlled in our model to avoid potential bias.

4 Lag and Decay of Impact

Unlike some consumer goods, where you may observe spike in sales shortly after running TV ads, it sometimes takes time for promotion impact to manifest in pharmaceutical marketing as patients need to make appointments with their physicians to discuss the treatment and patients and physicians may need repeated exposure to promotions to strengthen the impression and recall of drugs enough to take actions. As a result, the impact could be lagged and the product recall will decay over the time.

Properly identifying the delay and decay of promotion will be important to the accurate measurement of channel impact.

In the example below, changes in sales happen 1 week after the TV Ad broadcast. If a model tries to associate sales to GRPs in a non-lagged application, the results will return a negative relationship between sales and promotion, as sales increase when promotion volume is equal to zero. By lagging TV GRPs by 1 week, we can properly align the promotion spending with their impact on sales (Fig. 3).

We adopt an ad-stock approach to estimate the appropriate level of lag and decay for the effect of promotions: the effect of promotions extending several periods after the original delivery. Figure 4 illustrates that different channels may have dramatically different decay curves—in terms of curve shape or speed of decay.

To identify the most appropriate decay curves, we need to rely on the learning from our EDA analysis and use prior knowledge and data to identify the right decay curves.

After identifying the right lag and decay curves, we construct variable for each promotion channel capturing all the current and carry-over effect from historical promotions (Fig. 5).

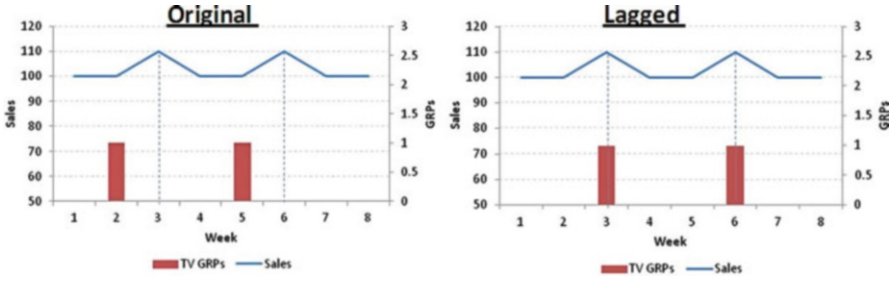


Fig. 3 Example of lagged effect of TV Ad

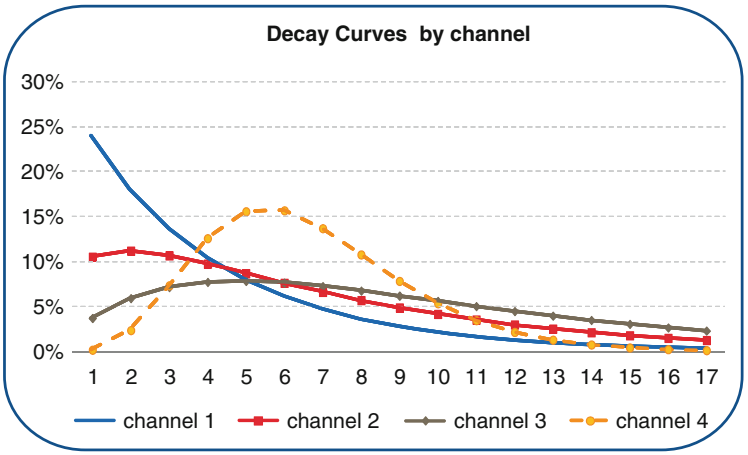


Fig. 4 Example of decay curves

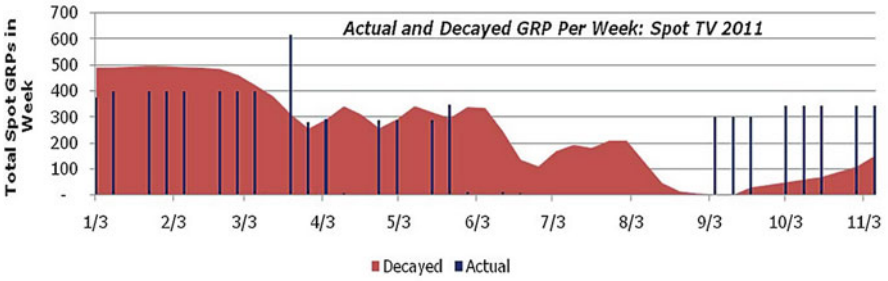


Fig. 5 Example of derived vs. original channel promotion

5 Model Structure

Each channel may have a different structure of impact on sales. For example, while samples may help physicians start trials with new patients, too many samples may lead to Rx substitutions or cannibalization, i.e., physicians may give out more samples than necessary for trial and end up with fewer prescriptions. TV ads, on the other hand, may not be effective at low spending level as it will be hard to achieve brand recognition and recall of ads among consumers without repeated exposures.

Figure 6 shows some common structures of impact. Samples could be structure 2 and TV ads should be structure 4.

After identifying the most possible structure(s) for each channel, we start to analyze impact by channel using the following promotion mix model: $y_{it} = a_0 + \mu_i + \sum_{k=1}^{n} b_k f_k(x_{kit}) + \varepsilon_{it}$, where y_{it} is the Rx of HCP i in week t, $\mu_i \sim N(0, \theta^2)$ is the

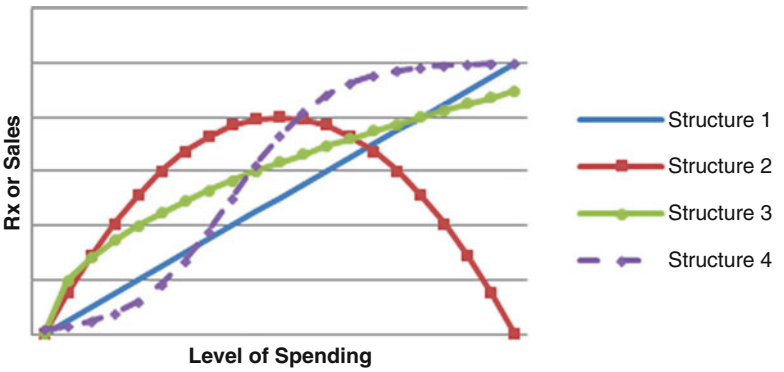


Fig. 6 Examples of impact structures

random effect at HCP level, $\varepsilon_{it} \sim N(0,\ \sigma^2)$ is the random error, $x_k,\ k = 1,\ 2 \ldots n$ represents the aligned/delayed spending of channel k, and $f_k()$ is the structure of that channel.

We normally try different model specifications and use out-of-sample validation in addition to standard model fit criteria, such as AIC, BIC to select the final model.

6 Business Application of Model Results

Using the model results, we can now try to answer the two business questions:

1. How effective are different channels in promoting sales?
2. How should we allocate resources across multiple channels?

To understand the first question, we need to look at the estimated model coefficients by channel and the corresponding response curves. The example in Table 1 shows that HCPs are responsive to this channel, as increased promotional level leads to higher sales, but the incremental return is diminishing as promotional level goes up. If the current level of promotion is 200 units per week and cost per unit is \$1, Return on Investment (ROI) = 480:280 = 1.71:1, i.e., on average, every \$1 invested in this channel brings \$1.71 back. If we increase level of promotion from 200 units per week to 250, we will be able to get additional \$70 sales back, or a net profit of \$70–\$50 = \$20, indicating that current spending level is too low. On the other side, if cost per Unit is \$2, we then over-spend in this channel, and cutting back would help raise the profit.

To answer the second question, we need to compare the response curves across multiple channels and take into consideration the total budget because, in most cases, marketers only have limited resources at disposal. Let's assume channel 1 has the response curve in Table 1, its cost per unit is \$1 and its current level of spending is 200 units per week. Now, assume the brand team has an additional \$50 to invest and a new channel becomes available to them whose response is listed in Table 2:

Table 1 Example of channel response

Unit of channel	Sales	Profit (cost per unit = $1)	Profit (cost per unit = $2)
0	0	0	0
50	150	100	50
100	280	180	80
150	390	240	90
200	480	280	80
250	550	300	50
300	600	300	0
350	630	280	−70
400	640	240	−160

Table 2 New channel response

Unit of channel	Sales	Profit (cost per unit = $2)
0	0	0
25	78	53
50	150	100
75	218	143
100	280	180
125	338	213
150	390	240
175	438	263
200	480	280

(1) If brand team chooses to invest in existing channel, the incremental sales would be $70 and the incremental profit will be $20 and (2) if brand team choose to invest in the new channel instead, the increase revenue would be $78 and the incremental profit will be $28. Obviously, to maximize the profit, brand team would choose to invest the incremental budget in the new channel.

7 Case Study

A major pharmaceutical company spends millions of dollars each year promoting one major product with annual sales close to $2 billion. The major promotion channels include: rep detailing, samples, TV Ads, newspaper Ads, professional journal ads, web banners, paid searches, etc. Table 3 shows a sample data that we collect for each HCP.

Next step is to identify the Lag/Decay of each channel and calculate ad-stock values for each channel. The example below (Fig. 7) shows that the correlation between brand new prescriptions (NRx) and local newspaper spending peaks at week 6. This is consistent with company's expectation as patients need time to make appointments with their physicians.

Table 3 Example data

Doc ID	Zip	DMA	Specialty	Week	Brand NRx	Market NRx	Details	Samples	Local News Paper	National TV GRPs
001253	60453	CHICAGO	PCP	4/16/2012	1	3.2	0	0	24.3	62.9
001253	60453	CHICAGO	PCP	4/23/2012	0.5	2.2	4	10	24.3	65.9
001253	60453	CHICAGO	PCP	4/30/2012	0	0	0	0	0	78.2
001253	60453	CHICAGO	PCP	5/7/2012	2	4.2	3	5	0	6.4
001253	60453	CHICAGO	PCP	5/14/2012	0	1	1	0	0	81.1
001253	60453	CHICAGO	PCP	5/21/2012	1	2.5	1	5	24.3	18.0

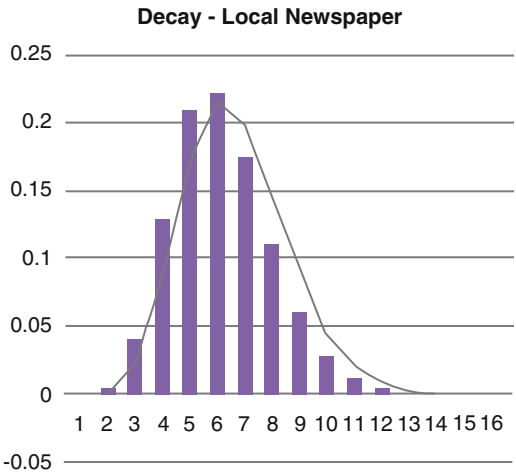


Fig. 7 Local newspaper impact decay curve

Table 4 Regression results (Partial)

Variable	Coefficient
Intercept	1.46
Details	0.82
Samples	1.12
HCP Journal	0.90
Newspaper	0.34
...	...

The selection of the proper functional form depends on the brand knowledge about the likelihood that marginal Rx return diminishes when spending increases. In this case, eventually a power additive model was selected: $\mathrm{NRx}_{it} = a_0 + \mu_i + \sum_{k=1}^{n} \mathrm{b}_k x_{kit}^{0.5} + \varepsilon_{it}$.

Part of the regression results are listed above (Table 4):

Take details as an example. The average cost per prescription is $80 and for every NRx, there are on average 1.3 refills. Thus 1 unit increase in NRx will lead

Table 5 Details response

Details	Incr. sales	Cost of details	Incr. profit
0	\$0	\$0	\$0
1	\$151	\$90	\$61
2	\$213	\$180	\$33
3	\$261	\$270	−\$9
4	\$302	\$360	−\$58
5	\$337	\$450	−\$113
6	\$370	\$540	−\$170

to incremental sales of $(1 + 1.3) \times \$80 = \$184$. The cost per detail is \$90. Thus we can create channel response for details using $\text{Incr.Sales} = 0.82 \times \text{Details}^{0.5}$.

In this example, we should invest no more than two details per week for a HCP to maximize brand profit (Table 5).

By comparing the responses across different channels, we can optimize the resource allocation.

8 Conclusions

Pharmaceutical companies heavily rely upon promotions to reach out to HCPs and patients to build brand recognition to gain market share and brand volume. Yet different channel's effectiveness could vary widely so promotion mix is important and it will influence the top line as well as the bottom line of a brand.

To understand the impact and contribution of each individual channel, we need to better understand the data we can access and gain insights about channel, such as lag and carry-over effects, by carefully studying the data.

Once we understand how each channel works, we can use promotion mix model to estimate channel impact and use that information to guide our future resource allocation to optimize sales.

References

The Congressional Budget Office (2009), Promotional Spending for Prescritpion Drugs

Duetsch LL (1998) Pharmaceuticals: The Critical Role of Innovation. In: Duetsch LL (ed), Industry Studies, Second Edition, M. E. Sharpe, New York

Grogan K (2010) More doctors closing the door on pharma sales reps, PharmaTimes RSS, 7 May 2010, Web

SK&A (2012), 2011 U.S. Pharma Company Promotion Spending, Feb 2012

Uplift Modeling Application and Methodology in Database Marketing

Junjun Yue

Abstract While there is a broad consensus that incrementality is the accurate measurement of database marketing impact, few marketing activities today are focused on uplift effect; because most of the target campaigns are selected by leveraging propensity models which maximize the gross response or demand. In this paper, we will introduce a tree-based uplift modeling methodology, which optimizes true marketing profitability. We will also discuss the major stages involved in this approach, with a real-life example from analytic services in the specialty retail industry.

1 Introduction

As a common practice in database marketing, we use response, conversion, and customer value models to optimize campaign response and revenue. These models predict the conditional probability or mean due to a market treatment, and they all fall under the general umbrella of propensity modeling—which is a statistical scorecard to predict the response likelihood or total expected purchase once a market treatment is given to the target audience. The treatment here could be a catalog mail, a discount offer, a creative message, or etc. Years of experience proved that these models worked well, achieving phenomenal return on investment (ROI) for database marketers through refined audience selection.

$$\text{Propensity Modeling} : F(\text{Response} \mid \text{Treatment})$$

F denotes the conditional probability or mean due to a market treatment

J. Yue (✉)
Merkle Inc., 7001 Columbia Gateway Drive, Columbia, MD 21046, USA
e-mail: Jyue@merkleinc.com

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_5,

However, media mix and multiple touches make these propensity models less and less efficient because a lot of target population no longer needs market treatments and they will respond any way. These "natural" responders come onboard due to established brand awareness and loyalty. This trend emphasized the importance of uplift modeling, which is to predict change in behavior from different treatments, e.g., mail vs. non-mail.

$$\text{Uplift Modeling} : F(\text{Response} \,|\, \text{Treatment A}) - F(\text{Response} | \text{Treatment B})$$

F denotes the conditional probability or mean due to a market treatment

2 Uplift Modeling Applications

Overall, there are two broad applications of uplift modeling in database marketing. One category is to identify customers with decent ROI via marketing treatment vs. non-treatment. For example, a catalog retailer may want to de-select customers with low or even negative purchase incremental from its catalog circulation; credit card companies will target credit line increase offers only to those customers who are likely to increase their spending and revolving dramatically which can offset the increased credit risk exposure. The other category is to segment customers based on different treatment preference. For example, an insurance company may want to improve quote conversions by allocating either a local agent or a remote dialer to contact a prospect based on predicted customer preference; a retailer will want to increase customer communication effectiveness by optimizing channel preferences (e.g., direct mail vs. email), offer preferences ($ Off vs. % Off; Product Focused vs. Discount Focused), frequency preferences (2 emails per week vs. 5 emails per week), etc.

3 Challenges of Uplift Modeling

Compared with traditional propensity models, there are four challenges for uplift modeling.

The first challenge is that one cannot pinpoint who provide incremental value at the individual level, the uplift impact of a specific marketing campaign has to be observed at population level. Otherwise, a traditional propensity model will suffice.

The second challenge is that uplift model will be much more unstable compared to gross propensity model. To illustrate this, let's assume a propensity model and we'll evaluate the model based on a 100 K random sample from the entire population. The top decile customers show a response rate of 5% and represent 10 K records which are 10% of the random sample. According to sample statistics and Central Limit Theorem, we can derive the actual response rate to fall in the range

of (4.6%, 5.4%) at 95% confidence level. The one-side deviation (0.4%) is only 8% of the observed response rate (5%). Now, let's look at an uplift model scenario and we'll need two additional inputs to get a feel of its volatility. One of the inputs is uplift percent due to the market treatment vs. non-treatment, in my retail CRM (customer relationship management) experience, 30% is a fair assumption. This is basically saying that the customers who receive direct marketing communication will have 30% higher response rate than those similar customers who are not communicated. The other input is treatment and non-treatment split in the sample data. Since uplift impact could only be observed at population level, a holdout (non-treatment) group is imperative. And we can assume 50–50 split which will mathematically minimize the volatility of uplift measurement. So in this case, we have 5 K top decile customers with treatment and show a response rate of 5%, and we have another 5 K top decile customers without treatment and show a response rate of 3.8% (5%/(1 + 30%)). Again according to sample statistics and Central Limit Theorem, the point estimate of uplift impact is 1.2% while the 95% confidence range is (0.4%, 2.0%). The one-side deviation (0.8%) is about 70% of the observed incremental response rate (1.2%).

The third challenge is that uplift modeling needs better planned sampling strategy compared to propensity model setup. Firstly, uplift modeling needs randomized treatment group as well as randomized holdout (or alternative treatment) group. Secondly, due to the much higher volatility encountered by uplift measurement, much larger sample size is needed to construct a robust uplift model compared with propensity models. In practice, we normally accumulate enough samples over multiple months even across a year instead of recommending an interruptive sampling in a short period of time.

The fourth challenge is that uplift modeling results often show counterintuitive results compared with propensity models. Most variables in final uplift model are very different compared with propensity models; and sometimes, the best customer segment from propensity models is the worst segment identified by uplift modeling since those folks are already over communicated or they have well established brand loyalty. The Pareto 20–80 principle suggests us to focus on the best customers, while the uplift modeling frequently suggests we should leave the best customers as they are and stop bothering them with marketing efforts. Although uplift modeling better optimizes true campaign impact, its high volatility (described as the second challenge) increases the likelihood of misclassification. In practice, we often overlay uplift model with propensity model to reduce the risk.

4 Uplift Modeling Methodology Review

In practice, there are four methods commonly used to investigate uplift effectiveness of direct marketing campaigns.

The first approach is two-model approach and there are three steps involved in this solution. Firstly, build a traditional propensity model on the sample records with

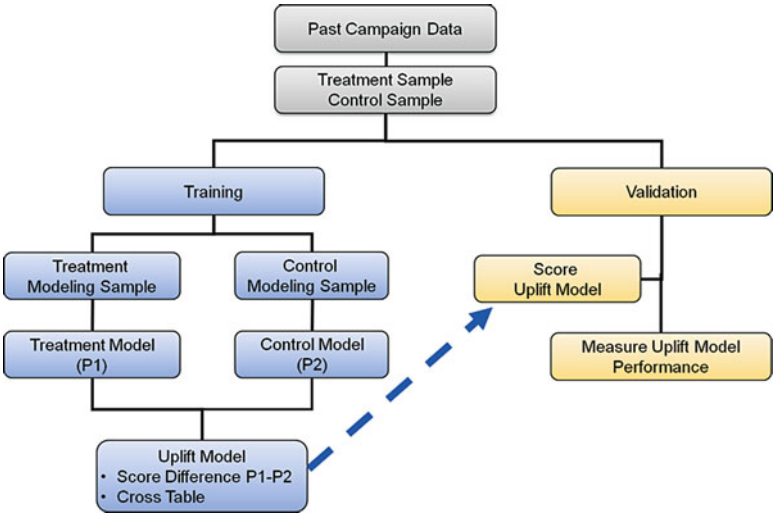


Illustration 1 Two-model approach

treatment; secondly, build a traditional propensity model on the sample records with no treatment or with alternative treatment; thirdly, develop uplift ranking criteria either by taking difference of the two model scores or by crossing tab the two model score segments. The advantage of two-model approach is that it leverages existing propensity model framework to the greatest extent so that it is simple to do. But the two-model approach tries to capture uplift impact through predicted model scores instead of modeling the actual uplift impact directly, and in only a few occasions, the two-model approach works as expected (Illustration 1).

The second approach is interaction modeling approach proposed by Lo (2002) and there are four steps involved in this solution. Firstly, combine treatment and non-treatment (or alternative treatment) observations into one dataset and create a treatment flag to indicate whether a treatment was applied to the customer or not, normally we set the flag to be 1 if the treatment was applied in the campaign and 0 if otherwise; Secondly, for all the independent variables $V_1, V_2, \ldots, V_N$ for modeling, we create interaction variables $IV_1, IV_2, \ldots, IV_N$ where $IV_N = \text{Flag} \times V_N$; thirdly, develop a model with intercept, main effects, treatment flag as well as interaction effects. For example, the model function form could be:

$$Y = a_0 + a_1 \times V_1 + a_2 \times V_2 + a_3 \times V_3 + b_0 \times \text{flag} + b_1 \times IV_1 + b_2 \times IV_2 + b_3 \times IV_3$$

Intercept—a_0, Main Effects—$a_1 \times V_1 + a_2 \times V_2 + a_3 \times V_3$, Treatment Flag—$b_0 \times$ flag, Interaction Effects—$b_1 \times IV_1 + b_2 \times IV_2 + b_3 \times IV_3$

And finally, derive uplift model function form from the overall model developed in step three by just keeping the model terms with treatment flag and interaction

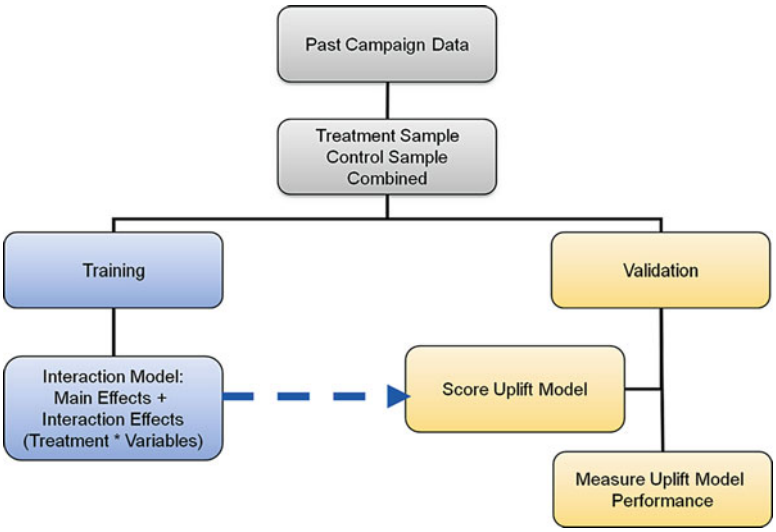


Illustration 2 Interaction modeling approach

effects while setting flag to be 1. For example, the final uplift model from the function form example above is:

$$Y = b_0 + b_1 \times V_1 + b_2 \times V_2 + b_3 \times V_3$$

The interaction modeling approach can theoretically optimize campaign uplift impact and will enable granular decision making in terms of uplift impact because each individual is scored with an uplift estimation. However, in practice, this approach often does not provide robust uplift results. This is possibly because the interaction variables selected from the modeling process are significantly correlated with gross response or demand, but not uplift impact (Illustration 2).

The third approach is a test design approach on predefined segments by leveraging business knowledge. For example, in order to investigate optimal email frequency, marketing practitioners often segment customers based on their historical RFM (recency, frequency, and monetary—how recent do they purchase, how frequent do they purchase, and how much do they purchase) as well as email open and click activities. And for each segment, we reserve enough sample size to observe the uplift impact of different treatments. As advantages, this approach utilizes existing business knowledge and frames uplift problems in a continuous test-learn agenda; also the sample size calculation through historical mean, historical standard deviation, practical significance, significance level, and power will guarantee us to have a significant uplift observation for predefined segments. But the predefined segments may not be the optimal segments to start with in terms of uplift impact and this approach does not have much flexibility to explore other dimensions which could be predictive of uplift impact; and the total test sample size needed increases quickly along with the expanded granularity of predefined segments.

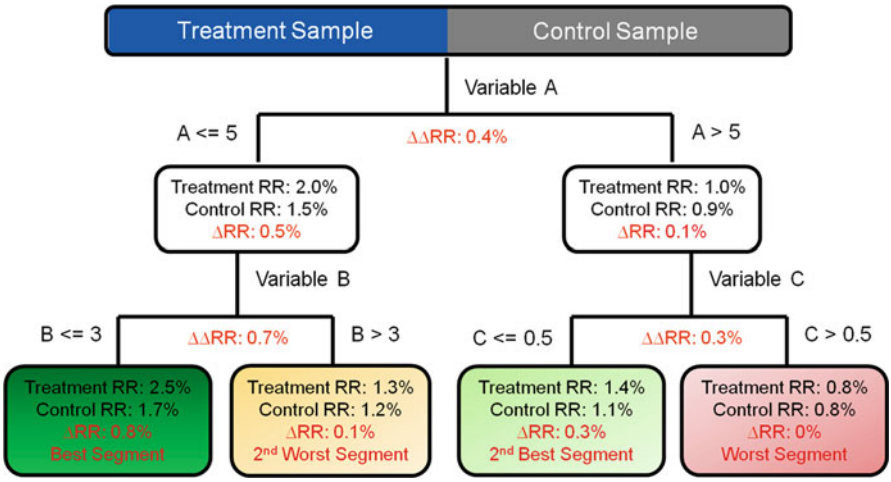


Illustration 3 Decision tree approach

The fourth approach is the decision tree approach which is aimed to identify the best variable as well as the best cutoff value of the variable for each binary split one at a time. Hansotia and Rukstales (2002) uses maximization of the difference in outcome between the two subpopulations (double difference) as the split criterion. Radcliffe and Surry (2011) proposes minimization of the difference in size between the two subpopulations as a second splitting criterion based on the facts that small segmented populations usually exhibit extreme outcome. The author developed a decision tree uplift modeling SAS application based on the criteria of (1) maximization of the difference in outcome between the two subpopulations; (2) pass of significance tests for each split; and (3) optional penalty for unbalanced split. And the author's methodology will be detailed in next section of the paper. On the positive side, the decision tree approach models uplift impact directly to drive optimization and the results are usually relatively stable after significance check and easy to explain. But on the negative side, this approach may not have enough sample size to observe uplift impact to very granular level; and it only optimizes one split at a time and does not consider possible interactions. Although decision tree approach is not a perfect solution for uplift modeling as well, it still seems to gain popularity in recent years (Illustration 3).

5 Tree-Based Uplift Modeling Methodology

To practically solve uplift related problems, the author developed a SAS application tool by building significance-based decision tree to optimize marketing uplift impact. The methodology is called SOS which stands for searching, optimization, and significance test.

5.1 Searching

This step is to go through all the variables as well as all the cutoff values of each variable based on specified range and step for each split. The purpose is to calculate the uplift statistics for the permutations of variables and variable cutoffs. For example, 300 variables with 5% step in the range of 20–80% result in $3,900 = (300 \times (80\,\% - 20\,\%/5\,\%) + 1))$ searching results. Once the optimal split of the parent node is determined, the tree will continue evolving to the next level based on recursive algorithm.

5.2 Optimization

This second step is to pick up the best split from all the possible permutations for each level of the tree. For each potential variable (X) and its possible cutoff value (A), the parent population will be split into two subpopulations—left branch ($X \leq \mathrm{A}$) and right branch ($X > A$). And within each branch, there are two subgroups—treatment A (e.g., mail) and treatment B (e.g., holdout). Uplift model works if and only if there is a significant difference between left branch uplift impact $\Delta \mathrm{RR_L} = \mathrm{RR_{L,A}} - \mathrm{RR_{L,B}}$ and right branch uplift impact $\Delta \mathrm{RR_R} = \mathrm{RR_{R,A}} - \mathrm{RR_{R,B}}$, this is to say, the absolute value of double difference $|\Delta \mathrm{RR_L} - \Delta \mathrm{RR_R}|$ should be significantly greater than 0. And among all the possible splits which result in significant double difference, the best split is to maximize double difference $|\Delta \mathrm{RR_L} - \Delta \mathrm{RR_R}|$, and with an optional subjection to the penalty of unbalanced split. The optional penalty factor the author applied in the tool is:

$$1 / \left(1 - f_{\mathrm{P}} \left| \frac{N_{\mathrm{L}} - N_{\mathrm{R}}}{N_{\mathrm{L}} + N_{\mathrm{R}}} \right|^{k} \right)$$

f_{P}: Penalty flag, 1 indicates application of unbalance split penalty; 0 indicates otherwise.
N_{L}: Total number of observations (treatment A + treatment B) in the left branch of the split.
N_{R}: Total number of observations (treatment A + treatment B) in the right branch of the split.
k: Penalty severity factor, $k \to 0$ indicates severer penalty of unbalanced split; $k \to +\infty$ indicates otherwise.

To allow more flexibility, the tool not only gives the best split information but also gives alternative significant splits based on optionally penalized double difference in a descending order. The reasons to provide alternative splits are:

- The best split may be purely derived from data mining and does not make business sense to marketers.
- The variable used as best split may not have good trend of uplift effect compared with other alternatives. For example, an upward or downward uplift sloping curve may be more ideal and robust for uplift modeling.

5.3 Significance Test

This third step is to use various significant tests to prune split alternatives so that (1) the best split automatically identified from the tool is robust to a great extent; (2) the tool will only end up with a manageable number of alternatives to be considered as uplift model splits.

The author incorporates two types of significance tests in the tool, they are (1) approximate U test; (2) contrast among treatments test.

5.3.1 Approximate U Test

According to Central Limit Theorem, the sample means $\mathrm{RR}_{\mathrm{L,A}}$, $\mathrm{RR}_{\mathrm{L,B}}$, $\mathrm{RR}_{\mathrm{R,A}}$, $\mathrm{RR}_{\mathrm{R,B}}$ (e.g., response rate, demand per circulation) in the uplift model are all random variables with approximate normal distribution because we normally have large enough (hundreds or thousands) observations in each of the four groups—left branch treatment A, left branch treatment B, right branch treatment A, and right branch treatment B. And since the individuals in each of the four groups are mutually independent, the four random variables $\mathrm{RR}_{\mathrm{L,A}}$, $\mathrm{RR}_{\mathrm{L,B}}$, $\mathrm{RR}_{\mathrm{R,A}}$, $\mathrm{RR}_{\mathrm{R,B}}$ are mutually independent as well. One tractable property of normal distribution is that: the linear combination of Independent random variables having a normal distribution also has a normal distribution. So the uplift measurement $(\mathrm{RR}_{\mathrm{L,A}} - \mathrm{RR}_{\mathrm{L,B}}) - (\mathrm{RR}_{\mathrm{R,A}} - \mathrm{RR}_{\mathrm{R,B}})$ turns out to be a random variable with approximate normal distribution. Therefore, we can apply a normal test statistics with mean $= (\mathrm{RR}_{\mathrm{L,A}} - \mathrm{RR}_{\mathrm{L,B}}) - (\mathrm{RR}_{\mathrm{R,A}} - \mathrm{RR}_{\mathrm{R,B}})$ and variance $= \frac{S^2_{\mathrm{L,A}}}{N_{\mathrm{L,A}}} + \frac{S^2_{\mathrm{L,B}}}{N_{\mathrm{L,B}}} + \frac{S^2_{\mathrm{R,A}}}{N_{\mathrm{R,A}}} + \frac{S^2_{\mathrm{R,B}}}{N_{\mathrm{R,B}}}$ to test the statistical significance of the double difference.

5.3.2 Contrast Among Treatments Test

The second significance test is derived from design of experiment. The uplift modeling is pretty much a two-factor, two-level DOE (design of experiment) framework. So the uplift problem could be rewritten as below:

$$\mu_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \varepsilon_{i,j,k} \, (i = 1, 2; \quad j = 1, 2)$$

$$\text{Hypothesis test}: \; H_0 : \gamma_{i,j} = 0; \quad H_1 : \gamma_{i,j} \neq 0$$

α_i: Main effect of treatment, β_j: Main effect of split, $\gamma_{i,j}$: Interaction effect of treatment and split

And a contrast in the context of design of experiment is defined as a special form of linear function of treatment group means, the mathematical expression is below:

$$C = \sum_{i=1}^{t} k_i \mu_i = k_1 \mu_1 + k_2 \mu_2 + \cdots + k_t \mu_t; \quad \text{where } \sum_{i=1}^{t} k_i = 0$$

μ_i: Sample mean of treatment group i, k_i: Numeric weight of treatment group i

Therefore, we can apply a t-test statistics with degree of freedom $= N - 4$ ($N = N_{\mathrm{L,A}} + N_{\mathrm{L,B}} + N_{\mathrm{R,A}} + N_{\mathrm{R,B}}$), mean $= \mathrm{RR}_{\mathrm{L,A}} - \mathrm{RR}_{\mathrm{L,B}} - \mathrm{RR}_{\mathrm{R,A}} + \mathrm{RR}_{\mathrm{R,B}}$, and variance $= \left(\frac{1}{N_{\mathrm{L,A}}} + \frac{1}{N_{\mathrm{L,B}}} + \frac{1}{N_{\mathrm{R,A}}} + \frac{1}{N_{\mathrm{R,B}}}\right) \times \mathrm{MSE}$ to test the statistical significance of the double difference.

And one interest relation between approximate U test and contrast among treatments test is that their mathematical expressions are the same if the uplift model sampling strategy is designed in a balanced way, i.e., treatment and non-treatment (or alternative treatment) have same number of observations. The proof is below:

$$\mathrm{MSE} = \frac{(N_{\mathrm{L,A}} - 1)S^2_{\mathrm{L,A}} + (N_{\mathrm{L,B}} - 1)S^2_{\mathrm{L,B}} + (N_{\mathrm{R,A}} - 1)S^2_{\mathrm{R,A}} + (N_{\mathrm{R,B}} - 1)S^2_{\mathrm{R,B}}}{N_{\mathrm{L,A}} + N_{\mathrm{L,B}} + N_{\mathrm{R,A}} + N_{\mathrm{R,B}} - 4}$$

$$N_{\mathrm{L,A}} = N_{\mathrm{L,B}} = N_1 \text{ and } N_1 \text{ is a large number}$$

$$N_{\mathrm{R,A}} = N_{\mathrm{R,B}} = N_2 \text{ and } N_2 \text{ is a large number}$$

$$\mathrm{MSE} \approx \frac{N_1 S^2_{\mathrm{L,A}} + N_1 S^2_{\mathrm{L,B}} + N_2 S^2_{\mathrm{R,A}} + N_2 S^2_{\mathrm{R,B}}}{2(N_1 + N_2)}$$

$$\mathrm{S_C} = \sqrt{\frac{S^2_{\mathrm{L,A}}}{N_{\mathrm{L,A}}} + \frac{S^2_{\mathrm{L,B}}}{N_{\mathrm{L,B}}} + \frac{S^2_{\mathrm{R,A}}}{N_{\mathrm{R,A}}} + \frac{S^2_{\mathrm{R,B}}}{N_{\mathrm{R,B}}}}, \text{ where } N_{\mathrm{L,A}} = N_{\mathrm{L,B}} = N_1 \textit{ and } N_{\mathrm{R,A}} = N_{\mathrm{R,B}} = N_2$$

6 Application of Tree-Based Uplift Modeling

One specialty retail client used campaign purchase models to make catalog circulation decisions. Although those propensity models ranked very well in terms of gross demand per circulation, we observed that the propensity models did not drive incremental purchases very effectively. A lot of top spending customers, especially those who are email-able and have fair amount of online web visiting and purchase activities, have very little uplift impact via monthly catalog communications. Considering the expensive catalog dollar per book, an uplift model is very appropriate to identify the low-incremental customer segments and cut back circulation accordingly. As a first step, we recommended to set up 20 K random mail as well as 20 K random control panels for 12 consecutive months (a balanced design). As mentioned in the beginning of the paper, uplift modeling is much more unstable compared with propensity modeling, and a large sample size is necessary. A white paper by Portrait Software (2006) pointed out that they would not expect the uplift modeling technique to work well on campaigns targeting fewer than 100,000 people. Once we accumulated the samples throughout the year, we divided the data

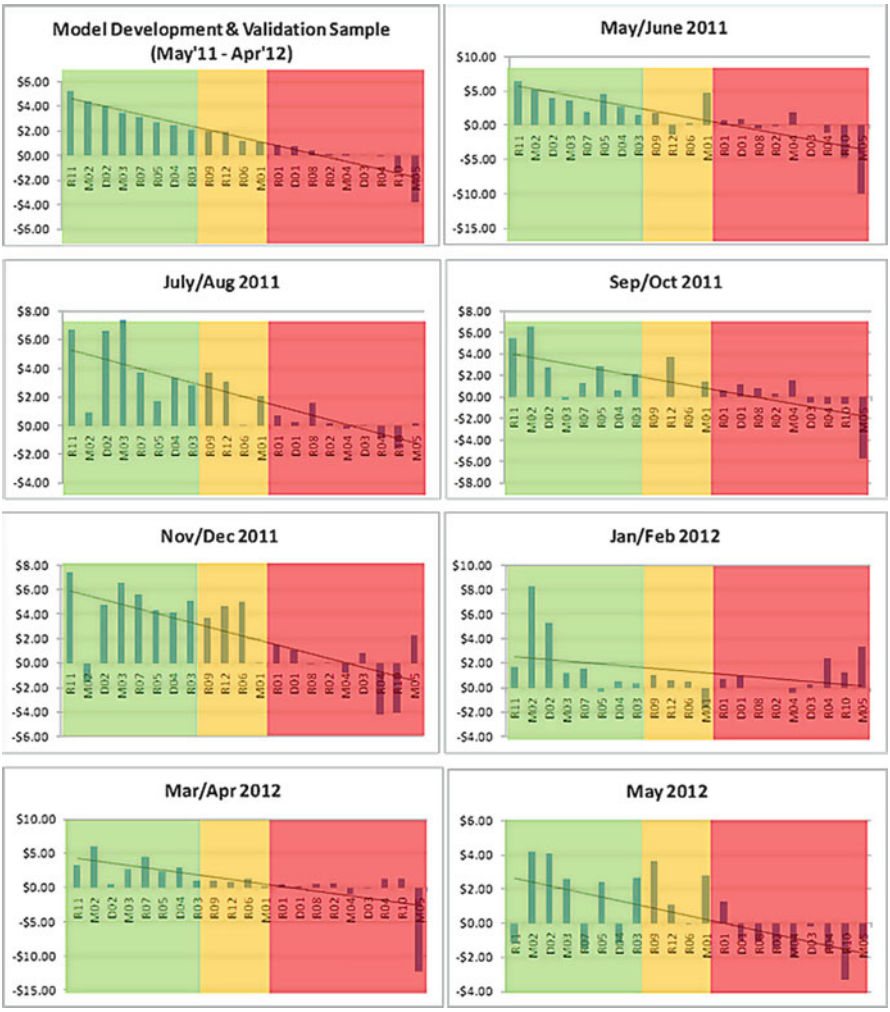


Illustration 4 Uplift model validation results

into 50% model development dataset and 50% out of sample validation dataset and applied the uplifting modeling SAS application tool to the development data to train the decision tree. Illustration 8 shows the uplift model validation results on the entire 12-month sample and samples of every two consecutive months. Although there are volatilities from month to month, the uplift model generally holds well over time. Through this initiative, we were able to identify ∼40% population with very minimum incremental who can be reduced from catalog circulation (Illustration 4).

We also set up random mail and control panels in future campaigns to back-test and verify uplift impact out of time across various mailing depths based on

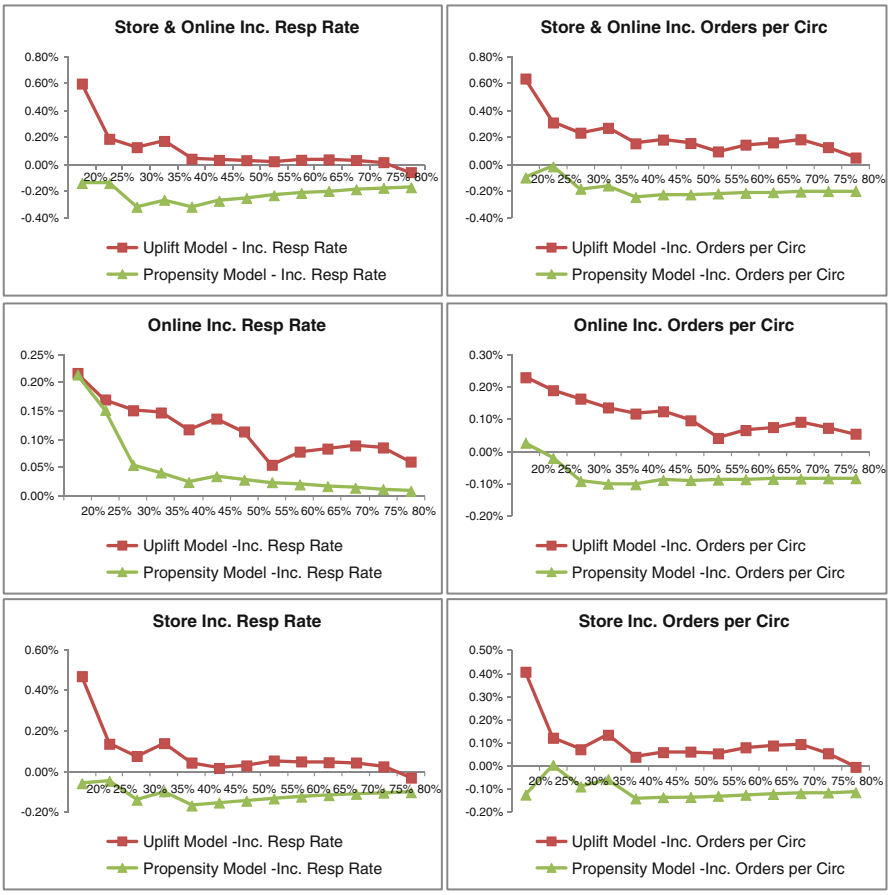


Illustration 5 Performance comparison—uplift model vs. propensity model

uplift modeling vs. traditional propensity modeling. The below results show that uplift modeling outperforms propensity modeling regardless of channels (store and online) and campaign metrics (response, orders) for this specialty retail client (Illustration 5).

7 Summary

Compared with traditional propensity models, the uplift modeling is a preferred methodology to maximize the true impact of marketing treatment. In this paper, the author discussed the challenges of uplift modeling, as well as the four frequently used approaches and their pros and cons.

Comparing with the other three approaches, significance-based decision tree is suggested by the author because (1) it models uplift impact directly; (2) it is relatively robust after applying significance test directly to uplift measurement (double difference); and (3) the results are easy to explain.

However, the uplift modeling has its rooted challenge of high volatility compared with propensity models. In practice, we sometimes overlay uplift model with propensity model together to reduce the risk of misclassification.

References

[1] Victor S. Y. Lo (2002). The true lift model—A Novel Data Mining Approach to Response Modeling in Database Marketing. *ACM SIGKDD Explorations Newsletter*, 4(2):78–86

[2] Nicholas J. Radcliffe and Patrick D. Surry (2011). Real-World Uplift Modeling with Significance-Based Uplift Trees. *Stochastic Solutions White Paper*

[3] Optimal Targeting through Uplift Modeling: Generating higher demand and increasing customer retention while reducing marketing costs. *A white paper by Portrait Software* (2006)

[4] Behram Hansotia and Brad Rukstales (2002). Incremental Value Modeling. *Journal of Interactive Marketing*, 16(3):35–46

Part II
Biomarker Analysis and Personalized Medicine

Designing Studies for Assessing Efficacy in Mixture Populations

Szu-Yu Tang, Eloise Kaizar, and Jason C. Hsu

Abstract In personalized medicine, the patient population is thought of as a mixture of two or more subgroups that might derive differential efficacy from a drug. A decision to make is which subgroup or union of the subgroups should the drug be developed for. Interestingly, some common measures of efficacy are such that its value for a mixture population may not be representable as a function of efficacy for the subgroups and their prevalence. This chapter describes *design* of study that would lead to probabilistic models so that relative risk (or odds ratio) for a mixture population can be represented as a function of relative risk (or odds ratio) for the subgroups and their prevalence.

1 Motivation

In personalized medicine, there is a question as to whether a drug should be marketed to the general population or only to a targeted subgroup. For example, consider a single biomarker which classifies patients into a g^+ subgroup and its complementary subgroup g^-. Drug developers might make their marketing decision based on the true efficacy of a treatment (vs. control) over the entire patient population, $\overline{\theta}$, and the efficacy within each subgroup, θ_{g^+}, θ_{g^-}.

Within-group efficacies are by definition conditional, where the effect size is conditional on group membership. Similarly, broad population efficacy is

S.-Y. Tang (✉) • E. Kaizar
The Ohio State University, Columbus, OH 43210, USA
e-mail: tang.142@buckeyemail.osu.edu; ekaizar@stat.osu.edu

J.C. Hsu
The Ohio State University, Columbus, OH 43210, USA

Eli Lilly and Company, Indianapolis, IN, USA
e-mail: jch@stat.osu.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_6,

unconditional, since no subgroup must be considered. However, to estimate broad efficacy, one might choose one of the two paths. First, one might choose an unconditional estimator, for example from a regression model that does not include a subgroup indicator. Alternatively, one might choose to use logical relationships to combine separate conditional estimates of the two subgroup efficacies. This latter approach might be carried out by combinations of parameter estimates from a regression model that includes the subgroup indicator.

This chapter discusses the two approaches to estimating broad population efficacy and their relationship with the study design. We first discuss how to derive an estimate of $\overline{\theta}$ by combining estimates of θ_{g^+} and θ_{g^-}. The methodology for computing such an overall estimate depends in part on the measure of efficacy. The choice of measure depends on both the nature of the disease being treated and the outcome of interest. For example, the primary outcome of a treatment for schizophrenia is usually the decrease in PANSS from baseline, and the efficacy of a drug is typically measured as the difference in the expected decrease, relative to a control treatment. However, a researcher would choose a different measure if he/she is interested in the safety of the new drug, where the outcome of interest is the binary experience of tremors (e.g., tardive dyskinesia) within 6 months of treatment. In this case, a more appropriate measure of efficacy may be the relative risk (RR) or odds ratio (OR) of tremor relative to a control treatment. We consider each of these measures of efficacy in turn.

2 Efficacy Measured as a Difference of Means

In this section we consider the difference of means as the measure of efficacy. Abbreviate "treatment" and "control" with *Rx* and *C*, respectively. Denote by μ^{Rx} and μ^{C} the mean responses over the entire patient population if the entire population had received treatment or control, respectively. Denote by $\mu^{Rx}_{g^+}$, $\mu^{Rx}_{g^-}$, $\mu^{C}_{g^+}$, $\mu^{C}_{g^-}$ the corresponding mean responses in the g^+ and g^- subgroups. We further consider the simple dataset presented in Table 1 to clarify the ideas presented in this section.

If a higher mean response is better, then broad population efficacy is defined:

$$\overline{\theta} = \mu^{Rx} - \mu^{C}.$$

Table 1 Data leading to different least squares means and marginal means

Genetic subgroup	g^+	g^-
Treatment (Rx)	3.5, 3.5, 3.5 (ave.=3.5)	7.5, 7.5 (ave.=7.5)
Control (C)	5, 6 (ave.=5.5)	4, 5, 6 (ave.=5)

Ignoring the subgroup designations, a natural estimate of broad population efficacy is the difference of sample means in the treatment and control groups. Applying this to our simple example gives:

$$\widehat{\overline{\theta}} = \hat{\mu}^{Rx} - \hat{\mu}^{C} = \frac{3.5+3.5+3.5+7.5+7.5}{5} - \frac{5+6+4+5+6}{5} = -0.1 < 0,$$

suggesting that the treatment is not effective for the broad population. This is the estimate achieved by using the MEANS statement in SAS linear modeling procedures. However, because the subgroup designation may be associated with treatment assignment, it is well known that this *marginal* or *unadjusted* approach may result in biased estimates. In particular, if the design is not balanced (i.e., uneven randomization within each subgroup), we expect the marginal estimator to be biased.

An alternative *adjusted* approach more appropriately combines the *conditional* estimates of efficacy for all of the subgroups. The effect of the drug in the g^+ and g^- subgroups are, respectively:

$$\theta_{g^+} = \mu^{Rx}_{g^+} - \mu^{C}_{g^+} \text{ and } \theta_{g^-} = \mu^{Rx}_{g^-} - \mu^{C}_{g^-}.$$

If the assignment to treatment/control is independent of subgroup, then we can write broad population efficacy as the weighted average of the subgroup efficacies:

$$\overline{\theta} = \mu^{Rx} - \mu^{C} = \gamma \times \theta_{g^+} + (1-\gamma) \times \theta_{g^-}, \tag{1}$$

where γ denotes the population prevalence of the g^+ subgroup.

Note that this relationship only holds for designs where the probability of randomization to treatment is constant across subgroups. This could be achieved either by an unstratified design (where the randomization probability must be independent of the unknown group membership) or in a stratified design with equal randomization probabilities across all strata. Following Good and Mittal (1987), we call this property of the study design *row uniformity in the parameters*.

Thus, for efficacy measured by the *difference* of mean treatment and control outcomes, the weighted average of the mean differences in the subgroups equals the expected difference in broad population under row-uniformity. We call the difference of means a measure of efficacy that is *mixture-representable*, because broad efficacy $\bar{\theta}$ can be represented as a *linear* function f of efficacy in the subgroups, and some known characteristic κ of the patient population,

$$\bar{\theta} = f(\theta_{g^+}, \theta_{g^-}, \kappa).$$

For the difference of averages Eq. (1), κ is the prevalence of the g^+ subgroup, denoted γ. Mixture-representable measures of efficacy are helpful for studies with enrollment stratified by biomarker (such as enrichment designs) and other designs that do not recruit and assign patients to treatment completely at random.

To estimate broad efficacy with a mixture-representable measure, an appropriate approach is to plug in known population values and estimated subgroup efficacies:

$$\widehat{\overline{\theta}} = f(\hat{\theta}_{g^+}, \hat{\theta}_{g^-}, \kappa),$$

where $\hat{\theta}_{g^+}$ and $\hat{\theta}_{g^-}$ are estimators of θ_{g^+} and θ_{g^-}. For the difference of means illustrated in our small dataset, if prevalence γ is known to be 50 %, then an appropriate estimator of $\overline{\theta}$ is:

$$\widehat{\overline{\theta}} = \gamma \times \hat{\theta}_{g^+} + (1-\gamma) \times \hat{\theta}_{g^-} = \frac{3.5+7.5}{2} - \frac{5.5+5}{2} = 0.25 > 0,$$

suggesting that the drug may be broadly effective. In linear models, this is referred to as the *least squares means* (LS-means) approach to estimation. In SAS, it can be executed using the ESTIMATE statement in linear modeling procedures.

Clearly the *adjusted* estimator differs from the *unadjusted* estimator because it takes the different sample proportions of treatment and control patients in the g^+ and g^- subgroups into account. That is, if there is any failure in the randomization that leads to uneven numbers of patients receiving treatment and control in the biomarker subgroups, the adjusted means approach still produces unbiased estimators.

However, if the design of the study has proportional cell frequencies across the subgroups (which we call *row-uniformity in the data*), then there is no need for adjustment. The intuition is that just as when Eq. (1) holds when the probability of assignment to treatment is constant across groups, the sample version of this equation holds when the sample proportion of those assigned to treatment is constant across groups. The proof comes from the fact that the information matrix of estimates of treatment and control effects for each subgroup is essentially a diagonal matrix with the proportional sample sizes as the diagonals, so least squares means are the marginal (unadjusted) means (see also Sect. 7.1.2 of Hsu 1996). Row-uniformity turns out to also be the key condition needed for the adjusted and unadjusted estimators of RR to coincide, as we show in the next section.

3 Efficacy Measured as a Relative Risk

For binary response (responder vs. non-responder, or adverse event vs. no adverse event), Good and Mittal (1987) gave a condition for the *design* of a study for RR to be mixture-representable. Let $p^{Rx}_{g^+(R)}$ represent the joint probability that a patient is assigned to treatment, belongs to the g^+ subgroup, and experiences a positive response. Let $p^{C}_{g^-(NR)}$ represent the corresponding joint probability for assignment to control.The analogous joint and marginal probabilities are shown in Table 2.

Table 2 Probabilities of response (responders (R) or non-responders (NR)), biomarker subgroup (g^+ or g^-), and treatment assignment (Rx or C). The internal cells of the two tables at the left display the three-way joint probabilities for the three classification variables. The marginal cells of these two tables represent the two-way joint probabilities for biomarker vs. the other variables. The table on the right displays the two-way and marginal probabilities ignoring the biomarker subgroups, so that the sum of the probabilities in corresponding cells of the two tables at the left equals the probability denoted in the corresponding cell of the right-hand table.

	g_+ subpopulation				g_- subpopulation				population		
	R	NR			R	NR			R	NR	
treatment (Rx)	$p^{Rx}_{g^+(R)}$	$p^{Rx}_{g^+(NR)}$	$p^{Rx}_{g^+}$	+	$p^{Rx}_{g^-(R)}$	$p^{Rx}_{g^-(NR)}$	$p^{Rx}_{g^-}$	=	$p^{Rx}_{(R)}$	$p^{Rx}_{(NR)}$	p^{Rx}
control (C)	$p^{C}_{g^+(R)}$	$p^{C}_{g^+(NR)}$	$p^{C}_{g^+}$		$p^{C}_{g^-(R)}$	$p^{C}_{g^-(NR)}$	$p^{C}_{g^-}$		$p^{C}_{(R)}$	$p^{C}_{(NR)}$	p^{C}
	$p_{g^+(R)}$	$p_{g^+(NR)}$	p_{g^+}		$p_{g^-(R)}$	$p_{g^-(NR)}$	p_{g^-}		$p_{(R)}$	$p_{(NR)}$	1

Let RR_+, RR_-, and $\overline{RR}$ denote RR for the g_+, g_- subgroup and broad population, respectively. Then these measures of efficacy are constructed by the probabilities:

$$RR_+ = \frac{\Pr(R|Rx, g^+)}{\Pr(R|C, g^+)} = \frac{p^{Rx}_{g^+(R)} \big/ p^{Rx}_{g^+}}{p^{C}_{g^+(R)} \big/ p^{C}_{g^+}} = \frac{p^{Rx}_{g^+(R)} p^{C}_{g^+}}{p^{C}_{g^+(R)} p^{Rx}_{g^+}} \tag{2}$$

$$RR_- = \frac{p^{Rx}_{g^-(R)} p^{C}_{g^-}}{p^{C}_{g^-(R)} p^{Rx}_{g^-}} \tag{3}$$

$$\overline{RR} = \frac{p^{Rx}_{(R)} p^{C}}{p^{C}_{(R)} p^{Rx}} \tag{4}$$

Using this notation, we call an experimental design *row uniform in the parameters* if the probability of assignment to treatment vs. control is constant across the subgroups and in the broad population, i.e.:

$$\frac{p^{Rx}_{g^+}}{p^{C}_{g^+}} = \frac{p^{Rx}_{g^-}}{p^{C}_{g^-}} = \frac{p^{Rx}}{p^{C}} = \lambda \tag{5}$$

In this case, the risk ratio is mixture-representable, where the mixing weight is the population proportion of control responders who are g^+, i.e., $\xi^+ = p^{C}_{g^+(R)} \big/ p^{C}_{(R)}$:

$$\overline{RR} = \xi^+ \times RR_+ + (1 - \xi^+) \times RR_- \tag{6}$$

The proof parallels Theorem 4.2 in Good and Mittal (1987). For a row-uniform design, the definitions of RR in Eqs. (2) and (3) simplify to:

$$RR_+ = \frac{1}{\lambda}\frac{p^{Rx}_{g^+(R)}}{p^{C}_{g^+(R)}} \text{ and } RR_- = \frac{1}{\lambda}\frac{p^{Rx}_{g^-(R)}}{p^{C}_{g^-(R)}}$$

Further, Eq. (4) becomes:

$$\begin{aligned}\overline{RR} &= \frac{1}{\lambda}\frac{p^{Rx}_{(R)}}{p^{C}_{(R)}} \\ &= \frac{1}{\lambda}\frac{p^{Rx}_{g^+(R)} + p^{Rx}_{g^-(R)}}{p^{C}_{g^+(R)} + p^{C}_{g^-(R)}} \\ &= \frac{1}{\lambda}\{\xi^+\lambda RR_+ + \left(1-\xi^+\right)\lambda RR_-\} \\ &= \xi^+ RR_+ + \left(1-\xi^+\right) RR_-\end{aligned}$$

Again, we have a choice of estimators for $\overline{RR}$. We could use either an unconditional or unadjusted estimator based on Eq. (4), or a conditional or adjusted estimator based on Eq. (6). If we choose the latter and ξ^+ is known, a natural adjusted estimator for broad RR is the weighted average of estimates $\widehat{RR}_+$ and $\widehat{RR}_-$ of the subgroup RRs:

$$\widehat{\overline{RR}} = \xi^+\widehat{RR}_+ + \left(1-\xi^+\right)\widehat{RR}_-$$

If the control treatment is standard of care, it may be reasonable to take the true population value of ξ^+ (the proportion of standard of care responders who are g^+) to be known with great precision, as estimated from historical data. However, even if much historical data is available, one should be cautious in considering a one-armed trial, as this would open the door to confounding. Our proposed *adjusted* estimator merely uses historical information to appropriately weight the two confounder-free conditional subgroup-specific estimates of RR. If a large volume of usual care data is not known or relevant, one may contemplate using a data-based estimate, $\hat{\xi}^+$, equal to the sample prevalence of the g^+ marker among control responders.

As for the difference of means, it is important to understand when *adjusted* estimators of RR are necessary, and when easier to compute *unadjusted* estimators give identical results. As we show below, it turns out that if the design of the study is *row uniform in the data*, *unadjusted* estimators coincide with their *adjusted* counterparts.

Let a_+, b_+ be the number of responders and non-responders for the treatment (R_x) arm, and let c_+, d_+ be the number of responders and non-responders for the control (C) arm in g_+ subgroup. Analogously, we define a_-, b_-, c_-, d_- for the g_- subgroup, as in Table 3. Then estimates of RR_+, RR_-, and $\overline{RR}$ are

Table 3 Number of responders (R) and Non-Responders (NR) for each genetic subgroup in the treatment and control arms and the combined group.

	g_+ subpopulation				g_- subpopulation				population		
	R	NR			R	NR			R	NR	
treatment (Rx)	a_+	b_+	N_+^{Rx}	+	a_-	b_-	N_-^{Rx}	$\Rightarrow$	$a_+ + a_-$	$b_+ + b_-$	N^{Rx}
control (C)	c_+	d_+	N_+^C		c_-	d_-	N_-^C		$c_+ + c_-$	$d_+ + d_-$	N^C
	R_+	NR_+	N_+		R_-	NR_-	N_-		R	NR	N

$$\widehat{RR}_+ = \frac{a_+}{N_+^{Rx}} / \frac{c_+}{N_+^C}$$

$$\widehat{RR}_- = \frac{a_-}{N_-^{Rx}} / \frac{c_-}{N_-^C}$$

$$\widehat{\widehat{RR}}_{\text{usual}} = \frac{a_+ + a_-}{N^{Rx}} / \frac{c + c_-}{N^C}$$

As noted in the previous section, an experimental design is said to be row uniform in the data if the ratio of the number of patients in the treatment arm to the number of patients in the control arm is the same for each of the genetic subgroups. That is, there exists a λ so that

$$\frac{N_+^{Rx}}{N_+^C} = \frac{N_-^{Rx}}{N_-^C} = \lambda \tag{7}$$

Thus, for a design row uniform in the data,

$$\widehat{RR}_+ = \frac{a_+}{N_+^{Rx}} / \frac{c_+}{N_+^C} = \frac{1}{\lambda} \frac{a_+}{c_+}$$

$$\widehat{RR}_- = \frac{a_-}{N_-^{Rx}} / \frac{c_-}{N_-^C} = \frac{1}{\lambda} \frac{a_-}{c_-}$$

Therefore, with $\hat{q} = \frac{c_+}{c_+ + c_-}$,

$$\begin{aligned}
\widehat{\widehat{RR}}_{\text{usual}} &= \frac{1}{\lambda} \frac{a_+ + a_-}{c_+ + c_-} \\
&= \frac{1}{\lambda} \{ \frac{c_+}{c_+ + c_-} \lambda \times \widehat{RR}_+ + \frac{c_-}{c_+ + c_-} \lambda \times \widehat{RR}_- \} \\
&= \frac{c_+}{c_+ + c_-} \widehat{RR}_+ + \frac{c_-}{c_+ + c_-} \widehat{RR}_- \\
&= \hat{q} \times \widehat{RR}_+ + (1 - \hat{q}) \times \widehat{RR}_- .
\end{aligned}$$

For any one realization of a randomized clinical study which is not stratified on genotype, it is unlikely that by chance the design will be exactly row-uniform

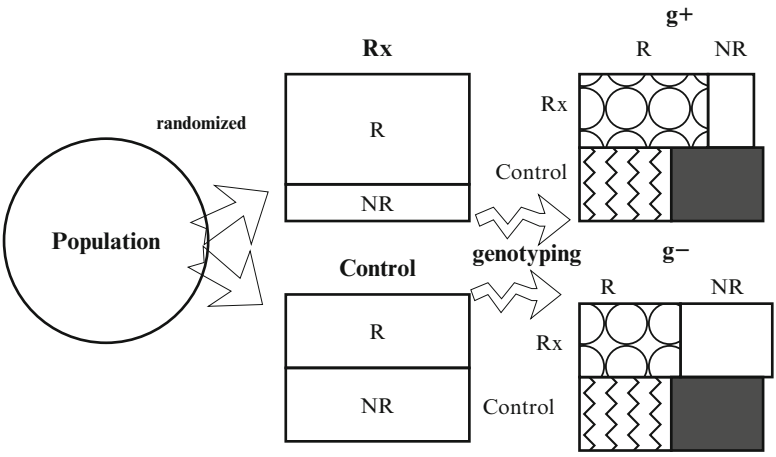


Fig. 1 Randomized design achieving approximate row-uniformity

in the data (i.e., the fraction of treated subjects in the g^+ group is identical to the corresponding fraction in the g^- group). In this case, the usual estimate of the broad risk ratio that ignores group membership is not equal to the weighted average of the individual risk ratios:

$$\widehat{RR}_{\text{usual}} \neq \hat{\xi}^+ \widehat{RR}_+ + \left(1 - \hat{\xi}^+\right) \widehat{RR}_-$$

Because this would be a desirable property (as for a balanced design in the difference of means case). In the next section we discuss possible designs that guarantee a row-uniform design in the data.

3.1 Designing Studies for RR to be Mixture-Representable

In a randomized clinical trial, randomizing the assignment of patients to treatment and control (Fig. 1) ensures that the treatment and control groups have similar characteristics on average, including row-uniformity in the data on average, i.e., the observed randomization fraction is on average the same in each biomarker subgroup. For large sample sizes, the law of large numbers suggests that most conducted studies will be close to row-uniformity in the data (e.g., Paré et al. 2010).

But in practice, it is possible for a randomized trial to lack sufficient row-uniformity in the data due to chance, especially if sample size is small. To avoid such a possibility, one can build exact row-uniformity into the study design by stratifying the study on genotype, and then randomizing the assignment of patients to treatment

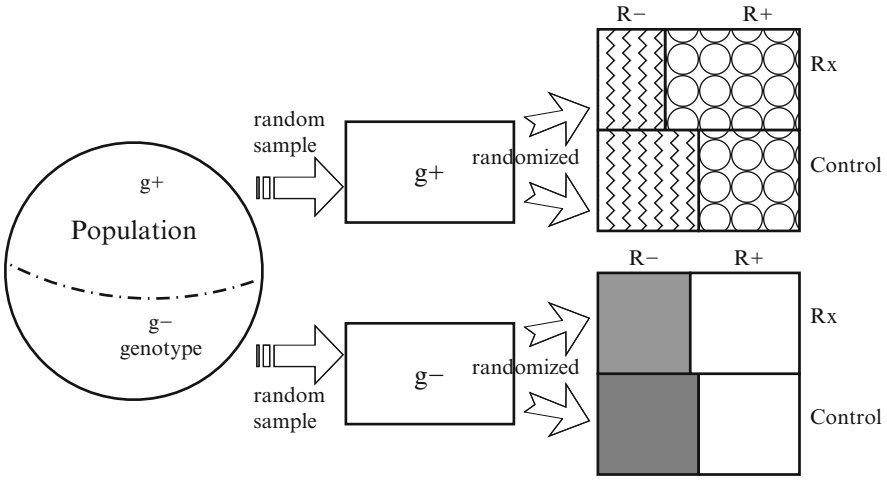


Fig. 2 Stratified design ensuring row-uniformity

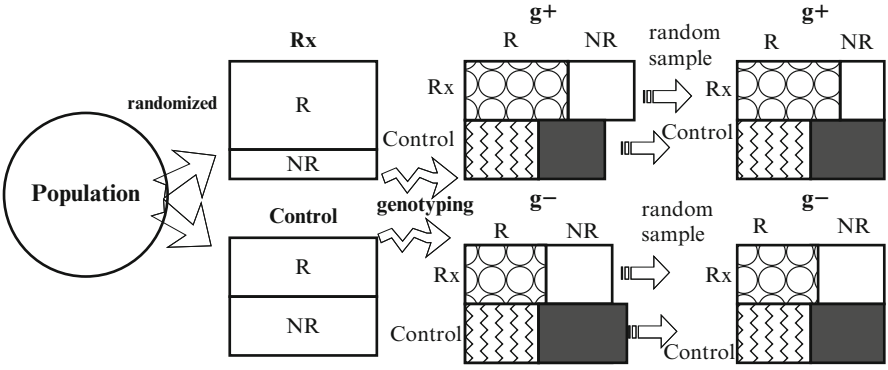


Fig. 3 Random sampling of stored samples to achieve row-uniformity

and control within the g^+ patients, and separately within the g^- patients, fixing the numbers of patients given treatment and control to achieve row uniformity in the data, as depicted in Fig. 2.

However, if patients have already been randomly assigned to treatment and control without stratification on genotype, then one can enforce exact row-uniformity in the data by randomly sampling the stored biological specimens of treated and control g^+ patients, and separately randomly sampling the stored biological specimens of treated and control g^- patients, with sample sizes that satisfy row uniformity in the data, as depicted in Fig. 3.

4 Odds Ratio as a Measure of Efficacy or Safety

The efficacy of a drug is sometimes quantified by the OR, the ratio of the odds of response between the treated and control groups. This measure is often convenient, as it is represented by the exponential of the parameters in standard logistic regression.

In safety studies, whether a genetic mutation is associated with patients experiencing an adverse event to a drug may be measured by the OR, ratio of the odds of patients experiencing an adverse event having this mutation, relative to patients without experiencing an adverse event having this mutation (e.g., Mallal et al. 2002).

Let OR_+, OR_-, and $\overline{OR}$ indicate OR for the g_+, g_- subgroup and for broad population, respectively.

An experimental design is said to be *column-uniform in the parameters* for g^+ and g^- subgroups if, for some η,

$$\frac{p_{g^+(R)}}{p_{g^+(NR)}} = \frac{p_{g^-(R)}}{p_{g^-(R)}} = \eta \tag{8}$$

So a design column-uniform in parameters is one in which the ratio of the probability of being a responder to the probability of being a non-responder is constant across the genetic subgroups (and therefore the combined population).

An experimental design is said to be *row-column-uniform* for g^+ and g^- subgroups if both conditions, Eqs. (5) and (8), are satisfied. The best known example of a row-column-uniform design *in data* is Fisher's "Lady Tasting Tea" experiment.

Theorem 4.3 of Good and Mittal (1987) shows OR is not subject to an amalgamation paradox for row-column-uniform designs. (An amalgamation paradox would occur when both OR_+ and OR_- are lower than $\overline{OR}$, or vice versa.) However, while row-uniformity is achievable as shown in the previous section, column-uniformity does not correspond to clinical expectation. Column-uniformity, constant ratio of or expected sample sizes of responders vs. non-responders across the genetic subgroup, corresponds to the biomarker being *independent* of response, $P(R|g^+) = P(R|g^-) = P(R)$. The more realistic expectation of a study with a thoughtfully chosen biomarker is depicted in Fig. 4, which is not column-uniform.

5 Concluding Remark

For continuous outcome measures, multiple comparisons such as Dunnett's method and Tukey's method were first implemented in SAS Proc GLM under the MEANS option, which estimates groups differences using marginal means unadjusted for possible different sample size proportions. However, by 1996, multiple comparisons were added to the LSMEANS option in Proc Mixed and Proc GLM, estimating

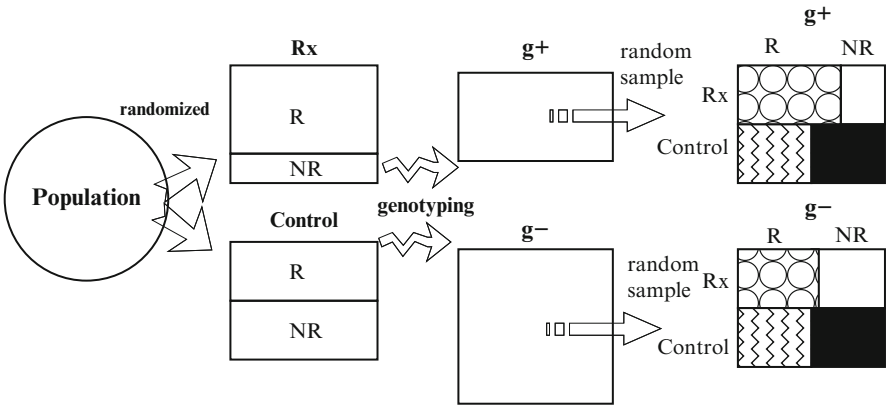


Fig. 4 Row-column-uniform Design I

group differences based on the least squares means approach. This chapter suggests similar caution may be needed when outcome measure is discrete.

Acknowledgments This research is supported in part by NSF grant DMS-1007794.

References

Good, I. J. and Mittal, Y. (1987). The amalgamation and geometry of two-by-two contingency tables. *Annals of Statistics*, 15:694–711.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall, London.

Mallal, S., Nolan, D., Witt, C., Masel, G., Martin, A. M., Moore, C., Sayer, D., Castley, A., Mamotte, C., Maxwell, D., James, I., and Christiansen, F. T. (2002). Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *The Lancet*, 359:727–732.

Paré, G., Mehta, S. R., Yusuf, S., Anand, S. S., Connolly, S. J., Hirsh, J., Simonsen, K., Bhatt, D. L., Fox, K. A., and Eikelboom, J. W. (2010). Effects of CYP2C19 genotype on outcomes of clopidogrel treatment. *New England Journal of Medicine*, 363:1704–1714.

Estimating Subject-Specific Treatment Differences for Risk-Benefit Assessment with Applications to Beta-Blocker Effectiveness Trials

Brian Claggett, Lu Tian, Lihui Zhao, Davide Castagno, and Lee-Jen Wei

Abstract In the recent past, several clinical trials have sought to evaluate the effectiveness of beta-blocking drugs in patients with chronic heart failure. Although the studies of certain drugs in this class yielded overwhelmingly positive results, other studies resulted in a much less clear interpretation. As result, attention has appropriately been placed on the impact of patient heterogeneity on treatment assessment. For clinical practice, it is desirable to identify subjects who would benefit from the new treatment from a risk-benefit perspective. In this paper, we investigate the results of the noted Beta-Blocker Evaluation of Survival Trial (BEST) and implement a systematic approach to achieve this goal by analyzing data available early in the study, at the time of a hypothetical initial interim analysis. We utilize multinomial outcome data from these initial patients to build a parametric score for the purpose of stratifying the remaining patients in the BEST study. We then use the data from the remaining BEST study participants to obtain a nonparametric estimate of the treatment effects, with respect to each of several ordered patient outcomes that encompass both risks and benefits of treatment, for any fixed score. Furthermore, confidence interval and band estimates are constructed to quantify the uncertainty of our inferences for the treatment differences over a

B. Claggett (✉) • L.-J. Wei
Harvard School of Public Health, Boston, MA 02115, USA
e-mail: bclagget@hsph.harvard.edu; wei@hsph.harvard.edu

L. Tian
Stanford University School of Medicine, Palo Alto, CA 94304, USA
e-mail: lutian@stanford.edu

L. Zhao
Northwestern University School of Medicine, Chicago, IL 60611, USA
e-mail: lihui.zhao@northwestern.edu

D. Castagno
University of Turin, Turin, Italy
e-mail: castagno.davide@gmail.com

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_7,

range of scores. We indeed detect subsets of patients who experience significant treatment benefits in addition to other patient groups who appear to be poor candidates for treatment.

1 Introduction

Consider a randomized, comparative clinical trial in which a treatment is assessed against a control with respect to their risk-benefit profiles. Conventionally, a single treatment contrast is utilized to assess an overall treatment difference with respect to efficacy, in addition to a global measure of toxicity, over a rather heterogeneous population. Unfortunately, the resulting inference about these two measures is rather difficult to interpret in clinical practice. Making patient-specific decisions based on estimated population-averaged effects can lead to suboptimal patient care (Kent and Hayward 2007). A positive (negative) trial based on these two overall measures does not mean that every future patient should (should not) be treated by the new treatment. Unfortunately, the typical ad hoc subgroup analysis of clinical studies is not credible (Wang et al. 2007). Moreover, such subgroup analysis is often conducted by investigating the effect of only a single predictor at a time and therefore may not be effective in identifying patients who would benefit from the new treatment. In this paper, we present a systematic approach to nonparametrically estimate subject-specific treatment differences from a risk-benefit perspective. To address the issue of risk-benefit assessment, we propose categorizing each patient into one of several ordered categories at a given follow-up time, more thoroughly reflecting the patient's experience during the study. In the event of censoring, we use inverse-probability weighting to obtain consistent estimates of the associated outcome probabilities. We use a training data set to first build a parametric univariate scoring system using baseline variables and then stratify subjects in the target data set accordingly. We then estimate the treatment effects nonparametrically with respect to the risk-benefit categories.

Data Our data set of interest comes from a clinical trial, "Beta-Blocker Evaluation of Survival Trial" (BEST), which compared a beta-blocker to placebo in patients with advanced chronic heart failure (HF), with a primary endpoint of all-cause mortality. In this trial, other monitored patient outcomes included timing of hospitalizations, with cause of each hospitalization recorded as being due to the patient's heart failure or for other reasons, and all deaths were adjudicated as being due to cardiovascular causes (CV death), or otherwise (non-CV death) (Beta-Blocker Evaluation of Survival Trial Investigators 2001). This trial enrolled 2,708 patients and is of interest because of the observed marginally significant treatment effect [HR= 0.90 (0.78, 1.02)] for the primary endpoint, standing in contrast to other similar studies of beta-blockers in patients with heart failure, each of which showed highly significant beneficial treatment effects (Domanski et al. 2003). The proposed training/validation approach allows one to avoid the nontrivial "self-serving" bias that can result from performing the model-building and variable selection process,

creating the score, stratifying subjects, and estimating subject-specific treatment differences all within the same data set. For this analysis, we mimic a hypothetical interim analysis, utilizing the first 900 patients ($\sim$33 % of total enrollment) in the trial as the training data for model building and selection, with the remaining 1,807 patients reserved to make inference on the effect of treatment over a range of the selected score values.

Patient Outcome Categories In order to both develop the scoring systems in the training data set and evaluate the patient outcomes in the validation set, a classification system is needed to describe each patient's status at a fixed follow-up time t_0. Our classification system involves five mutually exclusive and exhaustive categories, depending on which, if any, events have been experienced by a patient prior to t_0. This classification scheme is designed to account for events whose reduction represents an anticipated "benefit" of treatment (i.e., CV hospitalization, CV death), as well as events which may occur more frequently as a result of treatment, and are therefore considered potential treatment "risks" (i.e., non-CV hospitalization, non-CV death). For patient i receiving treatment j, let $T_{(1)ij}$ be a patient's time to first non-CV hospitalization, $T_{(2)ij}$ the time to the first CV-related hospitalization, and $T_{(3)ij}$ the time to death, $i = 1,...,n_j; j = 1,2$. For patients who are hospitalized for either non-CV or CV reasons, we note that $T_{(1)ij} < T_{(3)ij}$ and $T_{(2)ij} < T_{(3)ij}$, respectively. However, because death serves as a competing risk for hospitalization, a patient who dies without ever being hospitalized may be considered to have infinitely large values of $T_{(1)ij}$ and $T_{(2)ij}$. Let $T_j = \{(T_{(1)j}, T_{(2)j}, T_{(3)j}), j = 1,2\}$. Furthermore, let $\delta_{ij} = 1$ if a patient's death is classified as being CV-related, $\delta_{ij} = 0$ otherwise. In the following analysis, we investigate the joint distribution of $\{(T_j, \delta_j), j = 1,2\}$ for patients with given parametric score, which is needed for risk-benefit analysis on a personalized level. To this end, let us define a patient's classification at t_0 by ε_{ij}.

- $\varepsilon = 0$ if $t_0 < (T_{(1)} \wedge T_{(2)} \wedge T_{(3)})$ (patient is "alive and healthy")
- $\varepsilon = 1$ if $T_{(1)} \leq t_0 < (T_{(2)} \wedge T_{(3)})$ ("alive and hospitalized without worsening HF")
- $\varepsilon = 2$ if $T_{(2)} \leq t_0 < T_{(3)}$ ("alive with worsening HF")
- $\varepsilon = 3$ if $T_{(3)} \leq t_0, \delta = 0$ ("non-CV death")
- $\varepsilon = 4$ if $T_{(3)} \leq t_0, \delta = 1$ ("CV death")

2 Building a Scoring System Via the Training Data Set

To begin, we use the BEST training data set to build a scoring system using the patients' baseline characteristics with respect to the above ordinal outcome categories. Specifically, for this training set, each subject was assigned to a particular treatment j, where $j = 1$ denotes the untreated (placebo) group, and $j = 2$ denotes the treated (beta-blocker) group. Let U_j be the vector of baseline covariates, and let the treatment indicator τ_j with $\tau_j = 1$ in the treated group, and 0 otherwise.

Let C_j be the censoring variable, which is assumed to be independent of U_j and all T_j. Furthermore, let $X_{(1)j} = \min(T_{(1)j}, C_j), X_{(2)j} = \min(T_{(2)j}, C_j)$, and $X_{(3)j} = \min(T_{(3)j}, C_j)$ and $\{\Delta_{(r)j}, r = 1,2,3\}$ be the indicator function, which is one if $T_{(r)j} \leq C_j$. The data consist of $\{(X_{(1)ij}, X_{(2)ij}, X_{(3)ij}, \Delta_{(1)ij}, \Delta_{(2)ij}, \Delta_{(3)ij}, \delta_{ij}, U_{ij})', i = 1, \cdots, n_j\}$, n_j independent copies of $\{(X_{(1)j}, X_{(2)j}, X_{(3)j}, \Delta_{(1)j}, \Delta_{(2)j}, \Delta_{(3)j}, \delta_j, U_j)', j = 1,2\}$.

Now, suppose that we are interested in estimating the t_0-year outcome probabilities $\pi_{jk}(U), j = 1,2$, where $\pi_{jk}(U) = \text{pr}(\varepsilon_j = k|U)$ for a pre-specified time point t_0. To obtain estimates for $\pi_{jk}(U)$, one may use an ordinal regression working model of the following form

$$g(\gamma_{jk}(U_{ij})) = \alpha_k - \beta' Z_{ij} - \tau_{ij}(\beta^{*'} Z^*_{ij}) \tag{1}$$

where $\gamma_{jk} = \sum_{l=0}^{k} \pi_{jl}$ is the cumulative probability of a patient in treatment group j being classified into outcome category $\varepsilon \leq k$, Z_{ij} is a given function of U_{ij}, $Z^*_{ij} = (1, Z'_{ij})'$, $g(\cdot)$ is a given monotone function, and α, β_j and β^*_j are unknown vectors of parameters, with $\alpha = \{\alpha_k, k = 0, \ldots, 4\}$ representing the intercept terms and β_j and β^*_j corresponding to the main effects and treatment interaction effects, respectively, of the covariates Z on patient outcome status. Noting that a patient's outcome status is observable only when $\min(T_{(3)}, t_0) \leq C$, the parameter vectors above may be estimated by applying inverse probability of censoring weights and maximizing the standard weighted multinomial log-likelihood function

$$\sum_{ij} \frac{w_{ij}}{\hat{G}_j(X_{(3)ij} \wedge t_0)} \{\sum_{k=0}^{4} I(\varepsilon_{ij} = k) log(\pi_{jk}(U_{ij}))\}, \tag{2}$$

with respect to (α, β, β^*) where $\pi_{jk}(U_{ij}) = g^{-1}\{\alpha_k - \beta' Z_{ij} - \tau_{ij}(\beta^{*'} Z^*_{ij})\} - g^{-1}\{\alpha_{k-1} - \beta' Z_{ij} - \tau_{ij}(\beta^{*'} Z^*_{ij})\}$ via (1), $w_{ij} = I(X_{(3)ij} \leq t_0)\Delta_{ij} + I(X_{(3)ij} > t_0)$, $I(\cdot)$ is the indicator function, and $\hat{G}_j(\cdot)$ is the Kaplan–Meier estimator for $G_j(\cdot)$, the survival function of the censoring variable for the jth group, obtained by treating censored times as "events" and vice versa (Li et al. 2011; Uno et al. 2007; Zheng et al. 2006). Under some mild conditions, the resulting estimators $(\hat{\alpha}, \hat{\beta}, \hat{\beta}^*)$ converge to a finite constant vector as $n \to \infty$ even when the model (1) is not correctly specified (Uno et al. 2007). Note that one may repeatedly utilize (1) and (2) with various Z and $g(\cdot)$ via, for instance, a standard stepwise regression procedure with U, to obtain final estimates $\hat{\pi}_{jk}(U)$.

Evaluation of Working Models Since many variable selection procedures and link functions can be considered as candidates for estimating $\pi_{jk}(\cdot), j = 1,2$, it is important to formally evaluate their relative merits. To this end, we first note that the adequacy of such ordinal regression modeling procedures for t_0-year outcomes can be quantified by the cross-validated log-likelihood, where a larger log-likelihood suggests a better model fit and more accurate out-of-sample predictions. We use a repeated random cross-validation procedure, in each iteration randomly

Table 1 Candidate modeling procedures with average cross-validated log-likelihoods

Link Function	Variable Sel.	$\mathscr{L}$
logit	Full	-202.30
logit	AIC	-201.25
cloglog	Full	-201.18**
cloglog	AIC	-201.29

dividing the entire training data set into two mutually exclusive subsets, $\mathscr{B}$ and $\mathscr{H}$, with the "model building set" $\mathscr{B}$ comprising approximately 80 % of the full training set. For each model building set and for a given link function and variable selection procedure, we can construct a model, using only patients in $\mathscr{B}$ to estimate $\pi_{jk}(U)$, yielding predicted probabilities for the patients in the holdout set $\mathscr{H}$, with predictions given by $\hat{\pi}_{jk}(U_{ij})$. The cross-validated log-likelihood, adjusted for censoring, is $\sum_{(i,j)\in\mathscr{H}} \frac{w_{ij}}{\hat{G}_j(X_{(3)ij}\wedge t_0)}\{\sum_{k=0}^{4} I(\varepsilon_{ij}=k) log(\hat{\pi}_{jk}(U_{ij})\}$.

The modeling procedure yielding the largest cross-validated likelihood values will be used for the construction of our final working model and will be refit to the entire training data set in order to construct the final score. We used 16 clinically relevant covariates to fit the patient outcome data with various working models to estimate the probability of each outcome status at $t_0 = 18$ months. These baseline variables are: age, sex, left ventricular ejection fraction (LVEF), estimated glomerular filtration rate adjusted for body surface area (eGFR), systolic blood pressure (SBP), class of heart failure (Class III vs. Class IV), obesity (body mass index (BMI) > 30 vs. BMI ≤ 30), resting heart rate, smoking status (ever vs. never), history of hypertension, history of diabetes, ischemic heart failure etiology, presence of atrial fibrillation at baseline, and race (white vs. nonwhite). As in Castagno et al. (2010), we used three indicator variables to discretize eGFR values into four categories, with cut-points of 45, 60, and 75. We used the ordinal regression models described above to fit the outcome data and considered both the logit and complementary log–log links, $g(p) = log(\frac{p}{1-p}), g(p) = log(-log(1-p))$, respectively. For each of these two models, we used two different methods of variable selection. The first one used all 16 variables additively as well as their interactions with treatment. The second one used a stepwise regression procedure, stopping when no more covariates could be added/removed without subsequently increasing the Akaike information criterion (AIC).

To evaluate these modeling procedures, we used a repeated random cross validation procedure as described above. Even though the variables selected in each iteration of the cross validation procedure may be different, our goal is to assess the average performance of the given modeling procedures. In Table 1, we present these potential modeling procedures with their corresponding log-likelihood values, averaged over 50 cross-validation iterations. The full model using the complementary log-log link was found to provide the best overall fit in the cross-validation samples and is used derive our scores in the following section. The parameter estimates obtained from fitting this model to the full training data set are shown in Table 2.

Table 2 Ordinal regression coefficients for the final working model using BEST training data

	Full Model			Full Model	
Covariate	β	β^*	Covariate	β	β^*
Age	-0.001	-0.004	Never-smoker	0.126	-0.128
Sex: male	0.098	-0.148	Heart Rate	0.005	-0.015
LVEF	-0.014	-0.019	History of hypertension	0.213	-0.173
I(eGFR>75)	-0.175	-0.266	History of diabetes	0.308	-0.233
I(eGFR>60)	-0.041	-0.106	Ischemic etiology	0.083	0.103
I(eGFR>45)	-0.656	-0.078	Atrial Fibrillation	0.244	-0.286
SBP	-0.012	0.008	Race: white	0.044	-0.198
Class IV Heart Failure	0.191	0.634	I(BMI>30)	0.207	0.010
Treatment	-	1.268			

β represents main effects. β^* represents treatment interaction effects.

3 Potential Scoring Systems

Having completed this variable selection and model building step to select the "best" working model for predicting patient outcomes in both treatment groups, there are two reasonable ways in which to incorporate covariate information for the purposes of stratifying patients in the BEST trial.

Treatment Selection Score Perhaps the most commonly used method for stratifying patients is by estimated baseline risk, indexed here by $\beta' Z_{ij}$. Another interesting, though perhaps less commonly used, method for stratifying patients is according to treatment selection score (TSS), which is indexed by $-\beta^{*'} Z^*_{ij}$, the model-based estimate of the differential effect of treatment for a patient with given covariates. For a given U, let this score for the treatment contrast be denoted by $\hat{D}(U)$, which intends to estimate $D(U) = g(\gamma_{2k}(U)) - g(\gamma_{1k}(U))$ for any outcome category k if the modeling assumptions are true. Since γ_{jk} refers to the probability of being in a category equal to, *or healthier than*, category k, positive values of $D(U)$ correspond to an overall benefit associated with treatment. It is clear that if the modeling assumptions are valid, then this score directly addresses the question of whether or not a particular patient is a good candidate for treatment. This type of stratification system has recently been implemented successfully by Cai et al. (2011).

4 Making Inferences About the Treatment Differences over a Range of Scores with Respect to Ordered Patient Outcomes

Let the final parametric score for a patient with the covariate vector U in the target study be denoted by $S(U)$, which may be the risk score, based on the control group only, or the TSS $\hat{D}(U)$ discussed in the previous section. In order to make

inference about the risks and benefits of treatment at the patient level, we use the same ordered multinomial classification scheme described previously. We then construct the confidence interval and band estimates for the treatment differences with respect to the probability of a patient being classified into each possible clinical category. The target data consist of n_j independent and identically distributed observations as described in the training data. Furthermore, we let $\{Y_{ijk} = I(\varepsilon_{ij} = k), k = 0, \ldots, 4\}$, which is observable only if $\min(T_{(3)}, t_0) \leq C$. For the kth outcome, we are interested in estimating the treatment difference conditional on $S(U) = s$, that is, $E_k(s) = \text{pr}(\varepsilon_{i2} = k|S(U) = s) - \text{pr}(\varepsilon_{i1} = k|S(U) = s)$. To estimate $E_k(s)$ nonparametrically, we use a kernel estimator for each term on the right-hand side of the previous equation. Specifically, we estimate $p_{jk}(s) = \text{pr}(\varepsilon_{ij} = k|S(U) = s)$ with $\hat{p}_{jk}(s)$

$$= \left\{ \sum_i^{n_j} \frac{w_{ij} Y_{ijk}}{\hat{G}_j(X_{(3)ij} \wedge t_0)} K_{h_j}(V_{ij} - s) \right\} / \left\{ \sum_i^{n_j} \frac{w_{ij}}{\hat{G}_j(X_{(3)ij} \wedge t_0)} K_{h_j}(V_{ij} - s) \right\}, \qquad (3)$$

where $V_{ij} = S(U_{ij})$, $w_{ij} = I(X_{(3)ij} \leq t_0)\Delta_{(3)ij} + I(X_{(3)ij} > t_0)$, $\hat{G}_j(\cdot)$ is the Kaplan–Meier estimator of $G_j(\cdot)$, the survival distribution of the censoring variable C_j, estimated using observations $\{(X_{(3)ij}, \Delta_{(3)ij}), i = 1, \cdots, n_j\}$, $K_{h_j}(s) = K(s/h_j)/h_j$, $K(\cdot)$ is a smooth symmetric kernel with finite support and h_j is a smoothing parameter. Y_{ijk} may not be observable due to censoring, note that $w_{ij}Y_{ijk}$ is always observable. When $h_j = O(n^{-v}), 1/5 < v < 1/2$, it follows from a similar argument by Li et al. (2011) that $\hat{p}_{jk}(s)$ converges to $p_{jk}(s)$ uniformly over the interval $s \in \mathscr{S}$, where $\mathscr{S}$ is an interval contained properly in the support of $S(U)$. Let $\mathbf{E}(s) = \{E_0(s), \ldots, E_4(s)\}' = \mathbf{p}_2(s) - \mathbf{p}_1(s)$ and its empirical counterpart $\hat{\mathbf{E}}(s) = \{\hat{E}_0(s), \ldots, \hat{E}_4(s)\}' = \hat{\mathbf{p}}_2(s) - \hat{\mathbf{p}}_1(s)$, where $\hat{E}_k(s) = \hat{p}_{2k}(s) - \hat{p}_{1k}(s)$, $\mathbf{p}_j(s) = \{p_{j0}(s), \cdots, p_{j4}(s)\}'$, and $\hat{\mathbf{p}}_j(s) = \{\hat{p}_{j0}(s), \cdots, \hat{p}_{j4}(s)\}'$.

It follows from a similar argument by Li et al. (2011) that when h_j is of the same order as above, for a fixed s, the joint distribution $(n_1h_1 + n_2h_2)^{1/2}\{\hat{\mathbf{E}}(s) - \mathbf{E}(s)\}$ converges in distribution to a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\Sigma(s)$ as $n \to \infty$. To approximate this distribution, we use a perturbation-resampling method, similar to "wild bootstrapping" (Mammen 1993; Wu 1986) and has been successfully implemented in many settings (Cai et al. 2010; Lin et al. 1993; Park and Wei 2003). Details are provided in Appendix. To construct a $(1 - \alpha)$ simultaneous confidence band for $E_k(s)$ over the interval $\mathscr{S}$, we also use resampling methods, further described in Appendix. Here, we use an M-fold cross-validation procedure to choose the smoothing parameter $\hat{h}_j$ which maximizes a weighted cross-validated multinomial log-likelihood, as in Li et al. (2011). As in Li et al. (2011), $\hat{h}_j$ is of the order $n^{-1/5}$. To ensure the bias of the estimator is asymptotically negligible and that the above large-sample approximation is valid; however, we slightly undersmooth the data and let the final smoothing parameter be $\tilde{h}_j = \hat{h}_j \times n^{-0.05}$.

In order to aid in the interpretation of patient outcome probabilities, we additionally estimate patient-specific cumulative probabilities by repeating the same procedure as above using $\tilde{Y}_{ijk} = I(\varepsilon_{ij} \leq k), \gamma_{jk}(s) = E(\tilde{Y}_{ijk}|s)$, and $\Gamma_k(s) = \gamma_{2k}(s) - \gamma_{1k}(s)$. Each value $\Gamma_k(s)$ allows for the estimation of the treatment contrast with

respect to a different composite outcome. For example, $\Gamma_0(s)$ refers to the effect of treatment on the composite of "any hospitalization or death"; $\Gamma_2(s)$ corresponds to the effect of treatment on "any death," the initial primary outcome in the BEST study. While a particular component $E_k(s)$ may not always directly indicate whether a treatment is beneficial, the corresponding $\Gamma_k(s)$ will always have this desired interpretation, with positive values always indicating beneficial treatment effects.

5 Patient Outcomes in BEST Trial

Because covariates are expected to be balanced between the two treatment groups due to randomization, as an overall summary of the data set, we present for each treatment group the total number of patients in the target data set known to be in each outcome category, as well as the estimated cell probabilities, adjusting for censoring, with $P(\varepsilon_j = k)$ estimated via $\sum_i^{n_j} \frac{w_{ij}Y_{ijk}}{\hat{G}_j(X_{(3)ij} \wedge t_0)} / \sum_i^{n_j} \frac{w_{ij}}{\hat{G}_j(X_{(3)ij} \wedge t_0)}$. These results are shown in Table 3. We note that, overall, treated patients are somewhat more likely to be classified into outcome categories 0 and 1 (alive with no hospitalizations or non-CV hospitalizations only), and less likely to be classified into categories 2 and 4 (alive with CV-related hospitalization and CV death, respectively). The cumulative probabilities of a treated patient being classified at or below a certain threshold suggest a population-level beneficial effect of treatment, across all thresholds.

Now, we apply the final scoring system to the remaining 1,807 patients in the BEST trial, as derived from the full model described previously, with parameters estimated using the training set data. We note that 71 % of the target BEST patients are assigned scores greater than 0, indicating an anticipated treatment benefit for a majority of patients. To estimate $\mathbf{p}_1(s)$, $\mathbf{p}_2(s)$ and $\mathbf{E}(s)$ in our analysis below, we let $K(\cdot)$ be the standard Epanechnikov kernel. Using the TSS, we find that $\hat{D}$ is indeed related to the treatment effect on a patient's probability of experiencing outcome category $\varepsilon = 0$ (alive without any hospitalizations) and $\varepsilon = 2$ (alive with CV hospitalization), $\hat{E}_0(s)$ and $\hat{E}_2(s)$. Specifically, we find $\hat{E}_0(s) > 0$ for $s > 0.08$ and $\hat{E}_0(s) < 0$ otherwise. Patients with TSS $\in (0.30, 0.94)$ are found to be significantly

Table 3 BEST target data, 18-month patient outcomes: Observed patient outcomes and associated multinomial probability estimates, adjusted for censoring.

Outcome	**Control Group**			**Treated Group**		
Category	N	P($\varepsilon = k$)	P($\varepsilon \leq k$)	N	P($\varepsilon = k$)	P($\varepsilon \leq k$)
0	274	0.384	0.384	294	0.405	0.405
1	114	0.160	0.544	150	0.207	0.612
2	165	0.231	0.775	138	0.190	0.802
3	26	0.031	0.806	26	0.031	0.833
4	162	0.194	1.000	143	0.167	1.000
(censored)	156	-	-	159	-	-

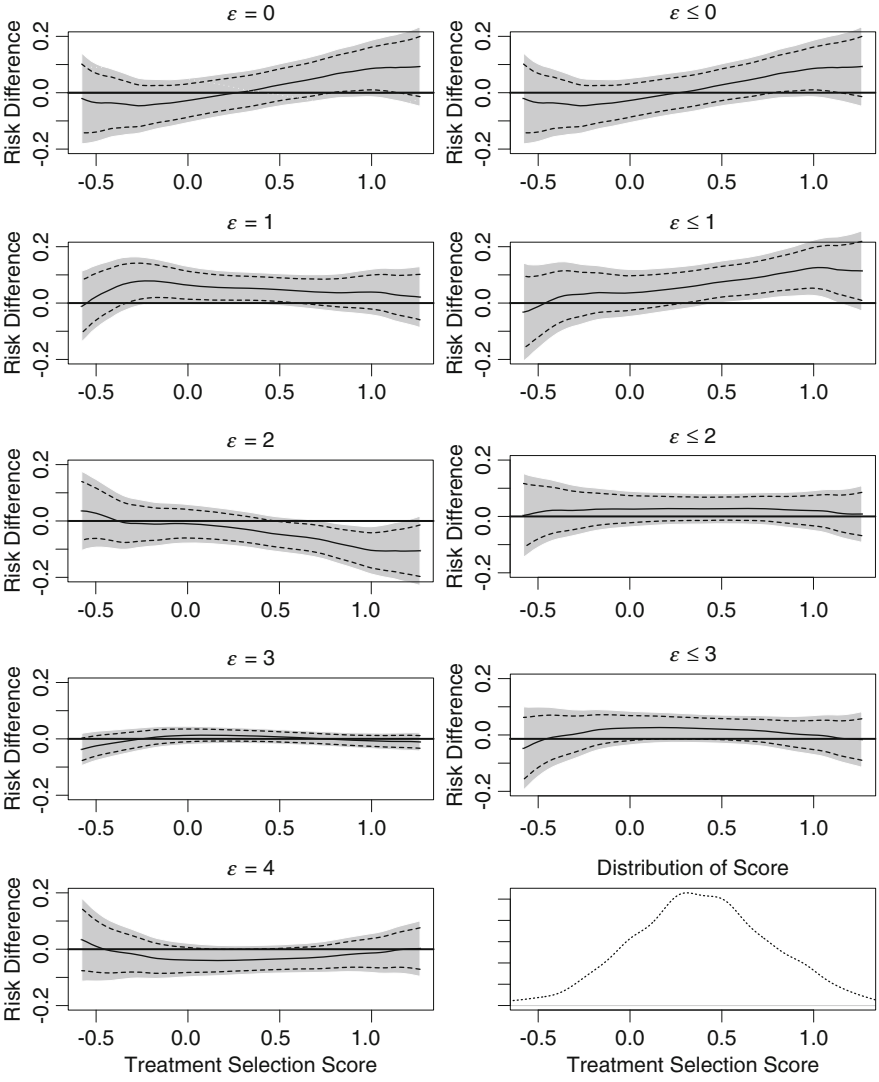


Fig. 1 BEST target treatment differences (treated minus untreated) using treatment selection score ($\hat{D}$) as a scoring system; Left panels: specific outcome probabilities; Right panels: cumulative outcome probabilities; Bottom right: Distribution of scores

more likely (via the 95 % confidence band) to experience outcomes $\varepsilon \leq 1$ (alive with no CV hospitalization), a range of scores representing approximately 36 % of the patients in the target data set. Treatment effects $\hat{\mathbf{E}}(\cdot)$ and $\hat{\mathbf{\Gamma}}(\cdot)$ are shown below in Fig. 1. It is interesting to note that the estimated (nonsignificant) effect of treatment in terms of reducing the risk of death is relatively constant, with a risk reduction of approximately 2 % across the range of scores.

Conclusions Ultimately, despite the nonsignificant overall result in the BEST trial, our scoring system is able to identify a sizeable subgroup of BEST patients who would experience significant benefits from treatment with the beta-blocker (i.e., bucindolol), in each case representing approximately one fourth to one half of the BEST patient population, depending on whether 95 % confidence intervals or bands are being used to determine statistical significance. The results from our analysis using the TSS are interesting for the reason that our scoring system, though not perfectly predictive of all clinical outcomes, seems to have done a reasonably good job of separating patients who would respond well to treatment from those who would respond poorly, indicating evidence of treatment interactions that are identifiable early in the course of the clinical trial and which are prospectively validated using future patients.

6 Remarks

In this paper, we use a two-stage process that can rather easily be applied in other scenarios with randomized clinical trial data. Here, we use two independent data sets obtained by separating the trial patients into two groups according to the order in which they were enrolled, thus the training data set is similar to that which may be available at the time of an interim analysis. We utilize these independent data sets to construct a systematic, subject-specific treatment evaluation procedure. The final scoring system may be chosen via a complex, exploratory model and variable selection process using the training data set. We then apply this system to stratify the patients in the second group and make inferences about the treatment effects with respect to patient outcomes for each stratum. If two similar studies are available (e.g., two Phase III trials are in many industry settings), one may instead use the first and second trials as the training and evaluation sets, respectively.

Because our goal is to develop a score to group patients with similar treatment responses, the TSS $\hat{D}(\cdot)$ should be the most effective system with which to stratify patients. However, there are practical concerns that could result in the preference of a risk score based on the control group only. In particular, if there are multiple treatments options to be compared, it will not be feasible to build a separate TSSs for each pairwise treatment comparison, and the baseline risk score would be an intuitive choice to investigate the effects of each treatment. A secondary analysis (not shown) of the current data set which utilized the baseline risk score suggested that low-risk patients from the BEST trial population may derive greater treatment benefits than high-risk patients.

In this paper, we used the t_0-year outcome probabilities as the outcomes of interest, where t_0 may be chosen from a clinical perspective. When there are two or more time points of interest, and particularly when the clinical endpoints being evaluated at these time points differ, it is unclear how to best summarize and analyze the resulting data. It would be appealing to choose a global measure to quantify the treatment contrast. Future research is needed in this area. Furthermore, our model

and variable selection procedure is intended to select the "best" model for fitting the data via the fitted log-likelihood. When the endpoint is the treatment difference, it is not clear that our approach would necessarily produce the "best" TSS. Further research is warranted along these lines.

Appendix

Construction of Confidence Intervals and Bands Let $\{B_{ij} : i = 1, ..., n_j; j = 1, 2\}$ be independent random samples from a strictly positive distribution with mean and variance equal to one. Let $p^*_{jk}(s)$ be the perturbed version of $\hat{p}_{jk}(s)$ with $p^*_{jk}(s) = \{\sum_i \frac{B_{ij}w_{ij}}{\widehat{G^*}_j(X_{(3)ij} \wedge t_0)} K_{h_j}(V_{ij} - s) Y_{ijk}\} / \{\sum_i \frac{B_{ij}w_{ij}}{\widehat{G^*}_j(X_{(3)ij} \wedge t_0)} K_{h_j}(V_{ij} - s)\}$. Here, $\hat{G}^*_j(\cdot)$ is the perturbed estimator for the survival function $G_j(\cdot)$ $\hat{G}^*_j(t) = \exp[-\sum_{i=1}^{n_j} \int_0^t \frac{B_{ij} d\{I(C_{ij} \le u \wedge X_{(3)ij})\}}{\sum_{l=1}^{n_j} B_{lj} I(X_{(3)lj} \ge u)}]$. Denote $\mathbf{E}^*(s) = \mathbf{p}^*_2(s) - \mathbf{p}^*_1(s)$, where $\mathbf{p}^*_j(s) = \{p^*_{j0}(s), \cdots, p^*_{j4}(s)\}'$. Using the arguments by Cai et al. (2010), the limiting distribution, conditional on the target data set, of $(n_1 h_1 + n_2 h_2)^{1/2}\{\mathbf{E}^*(s) - \hat{\mathbf{E}}(s)\}$, is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\Sigma(s)$. In order to obtain an approximation to $\Sigma(s)$, we generate a large number of realizations of $\{B_{i1}, B_{i2}\}$ from a standard exponential distribution, and compute $\mathbf{E}^*(s)$ for each perturbation sample. The resulting sample covariance matrix based on those perturbed estimates $\mathbf{E}^*$, say, $\tilde{\Sigma}(s)$, is a consistent estimator of $\Sigma(s)$. A two-sided confidence interval for an individual risk difference $E_k(s)$ is then given by $\hat{E}_k(s) \pm z_{(1-\alpha/2)}(n_1 h_1 + n_2 h_2)^{-1/2} \tilde{\sigma}_k(s)$, where $\tilde{\sigma}_k(s)$ is the kth diagonal element of $\tilde{\Sigma}(s)$.

To construct a $(1 - \alpha)$ simultaneous confidence band for $E_k(s)$ over the pre-specified interval $\mathscr{S}$, we cannot use the conventional method based on the sup-statistic, $\sup_{s \in \mathscr{S}} \tilde{\sigma}_k^{-1}(s) |(n_1 h_1 + n_2 h_2)^{1/2}\{\hat{E}_k(s) - E_k(s)\}|$ due to the fact that as a process in s, $(n_1 h_1 + n_2 h_2)^{1/2}\{\hat{E}_k(s) - E_k(s)\}$ does not converge to a process. On the other hand, one may utilize the strong approximation argument given in Bickel and Rosenblatt (1973) to show that an appropriately transformed sup of $\hat{E}_k(s) - E_k(s)$ converges to a proper random variable. In practice, to construct a confidence band, we can first find a critical value b_α such that $\text{pr}(\sup_{s \in \mathscr{S}} |E^*_k(s) - \hat{E}_k(s)| / \{(n_1 h_1 + n_2 h_2)^{-1/2} \tilde{\sigma}_k(s)\} > b_\alpha) \approx \alpha$. Then the confidence band for $E_k(s) : s \in \mathscr{S}$ is given by $\hat{E}_k(s) \pm b_\alpha (n_1 h_1 + n_2 h_2)^{-1/2} \tilde{\sigma}_k(s)$. Identical arguments are used for making inference with respect to $\Gamma_k(s)$.

Acknowledgments This manuscript was prepared using BEST Research Materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the BEST investigators or the NHLBI.

References

Beta-Blocker Evaluation of Survival Trial Investigators (2001). A trial of the beta-blocker bucindolol in patients with advanced chronic heart failure. *New England Journal of Medicine* **344,** 1659–1667.

Bickel, P. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics* pages 1071–1095.

Cai, T., Tian, L., Uno, H., Solomon, S., and Wei, L. (2010). Calibrating parametric subject-specific risk estimation. *Biometrika* **97,** 389–404.

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

Castagno, D., Jhund, P. S., McMurray, J. J., Lewsey, J. D., Erdmann, E., Zannad, F., ... & Dargie, H. J. et al. (2010). Improved survival with bisoprolol in patients with heart failure and renal impairment: an analysis of the cardiac insufficiency bisoprolol study II (CIBIS-II) trial. European journal of heart failure, **12**(6), 607–616.

Domanski, M., Krause-Steinrauf, H., Massie, B., Deedwania, P., Follmann, D., Kovar, D., Murray, D., Oren, R., Rosenberg, Y., Young, J., et al. (2003). A comparative analysis of the results from 4 trials of [beta]-blocker therapy for heart failure: Best, cibis-ii, merit-hf, and copernicus. *Journal of cardiac failure* **9,** 354–363.

Kent, D. and Hayward, R. (2007). Limitations of applying summary results of clinical trials to individual patients. *JAMA: the journal of the American Medical Association* **298,** 1209–1212.

Li, Y., Tian, L., and Wei, L. (2011). Estimating subject-specific dependent competing risk profile with censored event time observations. *Biometrics* **67,** 427–435.

Lin, D., Wei, L., and Ying, Z. (1993). Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika* **80,** 557–572.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics* **21,** 255–285.

Park, Y. and Wei, L. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika* **90,** 717–723.

Uno, H., Cai, T., Tian, L., and Wei, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102,** 527–537.

Wang, R., Lagakos, S., Ware, J., Hunter, D., and Drazen, J. (2007). Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine* **357,** 2189–2194.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics* **14,** 1261–1295.

Zheng, Y., Cai, T., and Feng, Z. (2006). Application of the time-dependent roc curves for prognostic accuracy with multiple biomarkers. *Biometrics* **62,** 279–287.

Missing Data in Principal Surrogacy Settings

Michael R. Elliott, Yun Li, and Jeremy M.G. Taylor

Abstract When an outcome of interest in a clinical trial is late-occurring or difficult to obtain, good surrogate markers can reliably extract information about the effect of the treatment on the outcome of interest. Surrogate measures are obtained post-randomization, and thus the surrogate–outcome relationship may be subject to unmeasured confounding. Thus Frangakis and Rubin (Biometrics 58:21–29, 2002) suggested assessing the causal effect of treatment within "principal strata" defined by the counterfactual joint distribution of the surrogate marker under the treatment arms. Li et al. (Biometrics 66:523–531, 2010) elaborated this suggestion for binary markers and outcomes, developing surrogacy measures that have causal interpretations and utilizing a Bayesian approach to accommodate non-identifiability in the model parameters. Here we extend this work to accommodate missing data under ignorable and non-ignorable settings, focusing on latent ignorability assumptions (Frangakis and Rubin, Biometrika 86:365–379, 1999; Peng et al., Biometrics 60:598–607, 2004; Taylor and Zhou, Biometrics 65:88–95, 2009). We also allow for the possibility that missingness has a counterfactual component, one that might differ between the treatment and control due to differential dropout, a feature that previous literature has not addressed.

1 Introduction

Given the time required to obtain clinical endpoints such as survival, there is interest in using surrogate endpoints such as disease-free survival at early follow-up periods [4], or biomarkers such as CD4 counts for AIDS [12] to assess the effectiveness of

M.R. Elliott (✉) • Y. Li • J.M.G. Taylor
Department of Biostatistics, University of Michigan School of Public Health, 1415 Washington Heights, Ann Arbor, MI 48105, USA
e-mail: mrelliot@umich.edu; yunsph@umich.edu; jmgt@umich.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_8,

a treatment regime in clinical trial settings. The demand for "surrogate markers" in clinical research has led to the development of a large number of statistical methods to evaluate the effectiveness of such measures [2]. Prentice's foundational paper defined "perfect" surrogacy as occurring when an outcome T is independent of treatment Z conditional on the surrogate measure S [15]. Buyse et al. [3] developed a meta-analytic approach that distinguishes between trial-level and individual-level surrogacy, with large coefficients of determination at both the individual and trial level indicative of a good surrogate measure.

An alternative approach to assessing surrogacy uses causal inference, with the goal of obtaining a surrogate whose causal treatment effects are strongly associated with the causal treatment effects on the outcome. Traditional regression models that condition on surrogacy measures to assess the fraction of the treatment effect explained can only be viewed as causal under the rather strong assumption that there are no confounders between the surrogate marker and the final outcomes, since the surrogate marker is observed post-randomization [17]. Robins and Greenland [16] define direct and indirect effects in mediation analysis in the potential outcomes framework. Assuming that the surrogate marker can be manipulated independently from the outcome, Robins and Greenland define (natural) direct effects as the expected value of the difference in the potential outcomes under different treatment assignments when the value of the marker is held constant, and indirect effects as the expected difference in the potential outcomes under treatment when the marker is changed to the value it would have taken under treatment and under control. An alternative "principal stratification" approach to assess surrogacy was proposed by Frangakis and Rubin [7]. Principal strata are defined by the joint potential outcomes of the surrogate marker, thus forming a "pre-randomization" variable that can be conditioned on while retaining causal interpretations of randomized treatment effects. The causal effects of interest become the differences in the potential outcomes under treatment and control within the strata in which the surrogate changes as a result of the treatment assignment. This approach has been explored for binary outcomes in more detail by Gilbert and Hudgens [8] and Li et al. [10]. Here we extend Li, Taylor, and Elliott [10] to accommodate missing data in the (observable) outcome measure, a common occurrence since the value of the surrogate variable is typically to provide information in advance of the outcome measure of interest. We utilize the machinery of the missing data literature, focusing on developing a non-ignorable missing mechanism for the final outcome of interest on the assigned treatment arm that is based on the assumption of latent ignorability [6, 14, 21]. We also allow for the possibility that missingness has a counterfactual component, one that might differ between the treatment and control due to differential dropout, a feature that previous literature has not addressed to our knowledge.

2 Assessing Surrogacy Via Principal Stratification

2.1 Notation

We denote treatment assignment by Z_l, the potential outcome for the surrogate under each of the treatment assignments for the lth subject by $S_l(Z_l)$, and the potential outcome for the true endpoint under each of the treatment assignments by $T_l(Z_l)$. We assume the surrogate is fully observed but the true endpoint under the assigned treatment arm is missing (for example, due to insufficient follow-up time or dropout), and denote $R_l(Z_l) = \{0,1\}$ corresponding to missing and observed true endpoints under each of the treatment assignments, respectively. Assuming dichotomous treatment assignments, surrogate markers, and true endpoints, the support for the joint distribution of the potential outcomes of the true endpoints is given by $\{(0,0),(0,1),(1,1),(1,0)\}$, corresponding, respectively, to failure under both arms, failure under control and success under treatment, success under both arms, and success under control and failure under treatment. Potential surrogate markers and responses have similar support corresponding to success/failure or observed/missing associated with each treatment arm. We denote the probability of a subject belonging to a cell in the resulting $4 \times 4 \times 4$ contingency table by $P((S(0),S(1)) = i,(T(0),T(1)) = j,(R(0),R(1)) = k) = \pi_{ijk}$, where $i,j,k = \{1,2,3,4\}$, corresponding to the four support points. Thus the "complete data" are $\{(S(0),S(1))_l,(T(0),T(1))_l,(R(0),R(1))_l; l = 1,\ldots,n\}$. The observed data for the lth subject is given by (z_l,r_l,s_l,t_l), where $r_l = R(Z_l = z_l)$, $s_l = S(Z_l = z_l)$, and $t_l = \begin{cases} T(Z_l = z_l) & \text{if } r_l = 1 \\ . & \text{if } r_l = 0 \end{cases}$, where "." indicates a missing value.

2.2 Surrogacy Measures of Interest

The principal strata correspond to the categories associated with the distribution of the potential surrogate markers, with (0,0) termed "never responsive", (0,1) "responsive", (1,1) "always responsive", and (1,0) "harmed" where without loss of generality 0 corresponds to a "poor health" surrogate marker and 1 to a "good health" surrogate marker. A common assumption is that the "harmed" stratum does not exist, often termed the "monotonicity" assumption since it implies $S_l(0) \leq S_l(1)$ for all l and thus $\pi_{4jk} = 0$ for all j,k; similarly monotonicity for the outcome implies $T_l(0) \leq T_l(1)$ and thus $\pi_{i4k} = 0$ for all i,k. A more limited form of monotonicity, which we term "stochastic monotonicity," only assumes the treatment is more likely to be helpful than harmful ($\pi_{2++} > \pi_{4++}$) for the surrogate measures, and that, within the non-harmful principal strata, the treatment is more likely to be helpful than harmful ($\pi_{j2+} > \pi_{j4+}$, $j = 1,2,3$) for the final outcome [5]. Under the assumption of "monotonicity," corresponding to no one harmed with respect to either the surrogate or outcome, the overall causal effect of treatment (CE) is given

by $E(T_l(1) - T_l(0)) = \pi_{+2} = \pi_{12} + \pi_{22} + \pi_{32}$. (We denote $\pi_{ij} \equiv \pi_{ij+} = \sum_k \pi_{ijk}$, the joint distribution of the potential surrogate marker and potential outcome marginalized across the missingness patterns.) Frangakis and Rubin [7] proposed associative and dissociative effects corresponding, respectively, to the fraction of patients on which the treatment improved both the surrogate marker and the final outcome AE=π_{22} and the fraction of patients on which the treatment changed the final outcome but not the surrogate marker DE=$\pi_{12} + \pi_{32}$. A good surrogate will have a large AE, indicating that the causal effect on the surrogate is highly associated with a causal effect on the final outcome of interest. Similarly, a good surrogate will have a small DE, indicating that the causal effect on the outcome is small when the causal effect on the surrogate is zero. Because AE and DE are constrained to sum to CE, interpreting "large" and "small" is perhaps most useful relative to the CE; hence Taylor et al. [20] defined associative proportions AP=AE/CE and dissociative proportions DP=DE/CE as the fraction of the overall treatment effect partitioned between the associative and dissociative effects. Frangakis and Rubin [7] suggest the concept of "perfect surrogacy," defined as causal effects on T occurring only with causal effects on S (i.e., $\pi_{12} = \pi_{32} = 0$); Li, Taylor, and Elliott [10] extend this to propose another surrogacy measure, CAP, which assesses the degree to which causal effects on T occur only with causal effects on S simultaneously with causal effects on S occurring only with causal effects on T: CAP=$\frac{\pi_{22}}{\pi_{12}+\pi_{21}+\pi_{22}+\pi_{23}+\pi_{32}}$. Without the monotonicity assumption, CE=$\pi_{+2} - \pi_{+4}$ (the net treatment effect for the final outcome corresponding to the fraction responsive to the treatment minus the fraction harmed), AE=$\pi_{22} + \pi_{42} - (\pi_{24} + \pi_{44})$ (net treatment effect on patients whose surrogate was responsive to treatment), and DE=$\pi_{12} + \pi_{32} - (\pi_{14} + \pi_{34})$ (net treatment effect on patients whose surrogate was not responsive to treatment) [11], and AP and DP are unchanged. Note that, in the absence of monotonicity, there is the possibility that treatment effects may be positive in some principal strata and negative in others, allowing the values of AP and DP to range outside of 0 and 1. In this case, the proposed marker may no longer be a reliable surrogate.

3 Principal Stratification Model for Surrogacy Accounting for Nonresponse in the True Endpoint

3.1 Model Assumptions

Factoring the joint distribution of the complete data we obtain

$$p(T(Z), S(Z), R(Z) \mid Z)) = p(T(Z) \mid S(Z), R(Z), Z))p(S(Z), R(Z), \mid Z))p(R(Z) \mid Z))$$

for $T(Z) = (T_1(Z), \ldots, T_n(Z))$, where $T_l(Z)$ refers to the set of potential outcomes for the lth subject associated with all possible treatment assignments Z in the sample, and similarly for $S(Z)$ and $R(Z)$.

We make the following three assumptions:

1. *Randomization:* Treatment assignment is made independently of the potential outcomes for the surrogate markers, so that

$$p(T(Z),S(Z),R(Z) \mid Z)) = p(T(Z),S(Z),R(Z))$$

2. *Stable unit treatment assignment [18]:* Treatment assignment for subject i is independent of $(S_j(Z_j),T_j(Z_j),R_j(Z_j))$ for $j \neq i$, so that

$$p(T(Z),S(Z),R(Z)) = \prod_l p(T_l(Z_l),S_l(Z_l),R_l(Z_l));$$

also the observed surrogate marker is equal to the potential outcome under the observed treatment arm ($s_l = z_l S(z_l) + (1-z_l)S(z_l)$), and similarly for T_l.

3. *Latent ignorability of missing data [6, 14, 21]:* Conditional on the joint distribution of the surrogate markers under both treatment assignments $S_l(Z_l)$, the joint distribution of the true endpoint under both treatment assignments $T_l(Z_l)$ is independent of counterfactual response. Thus we have

$$p(T_l(Z_l),S_l(Z_l),R_l(Z_l)) = p(T_l(Z_l) \mid S_l(Z_l),R_l(Z_l))p(S_l(Z_l),R_l(Z_l)) =$$

$$p(T_l(Z_l) \mid S_l(Z_l))p(S_l(Z_l),R_l(Z_l)).$$

We also consider a missing completely at random (MCAR) model in which missingness is independent of both the surrogate and the true outcome:

$$p(T_l(Z_l),S_l(Z_l),R_l(Z_l)) = p(T_l(Z_l) \mid S_l(Z_l))p(S_l(Z_l))p(R_l(Z_l))$$

Note also that we do *not* make the compound exclusion restriction (CER). Under CER, $S_l(0) = S_l(1)$ implies that $R_l(0) = R_l(1)$ and $T_l(0) = T_l(1)$ [6, 14]. Doing so would imply that subjects who had no causal impact of treatment on the surrogate outcome would have no causal impact of treatment either on the true endpoint of interest or on their response behavior, thereby assuming away key issue that we would like the data to speak to in our analysis.

3.2 Model Estimation

Factoring $P((S(0),S(1)) = i,(T(0),T(1)) = j,(R(0),R(1)) = k) = \pi_{ijk}$ as

$$P((T(0),T(1)) = j \mid (S(0),S(1)) = i,(R(0),R(1)) = k)$$
$$P((S(0),S(1)) = i,(R(0),R(1)) = k) = \pi_{j|ik}\pi_{i+k}$$

we have under the latent ignorability assumptions that $\pi_{j|ik} \equiv \pi_{j|i}$ for all k. We have 27 free parameters in the model (19 under monotonicity) but only 10 sufficient statistics in the data. Hence we use a fully Bayesian approach to cope with the non-identifiability in the observed data likelihood [9].

The complete data likelihood is given by

$$\prod_i \prod_j \prod_k \pi_{ijk}^{n_{ijk}} = \prod_i \prod_j \pi_{j|i}^{n_{ij+}} \prod_i \prod_k \pi_{i+k}^{n_{i+k}}$$

where n_{ijk} denotes the complete-data cell counts. (Note that the observed data cell counts are linear combinations of these complete-data cell counts; for example, the cell count of subjects for whom both the surrogate and outcome are observed to be failures under control (i.e., $z_l = 0, r_l = 1, s_l = 0, t_l = 0$) is given by $\sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=3}^{4} n_{ijk}$.) We assume a Dirichlet prior for the cell probabilities:

$$p(\pi_{j|i}) \sim DIR(a_{j|i})$$
$$p(\pi_{i+k}) \sim DIR(b_{i+k})$$

Under the MCAR assumption, $P((S(0),S(1)) = i,(R(0),R(1)) = k) = P((S(0),S(1)) = i)P((R(0),R(1)) = k)$ and thus $\pi_{ijk} = \pi_{j|i}\pi_{i++}\pi_{++k} = \pi_{ij+}\pi_{++k}$. The distribution of R is thus independent of S and T and thus can be ignored in both the data augmentation step and the draw of the parameters conditional on the complete data for estimation of the surrogacy effects of interest (although required to obtain model fit estimates discussed in Sect. 2.2). With 18 free parameters (11 under monotonicity) and 8 sufficient statistics, we again use a fully Bayesian approach, assuming a Dirichlet prior for the cell probabilities:

$$p(\pi_{ij+}) \sim DIR(a_{ij+})$$
$$p(\pi_{++k}) \sim DIR(b_{++k})$$

Estimation proceeds via Gibbs sampling and data augmentation: draws of the cell probabilities are made conditional on the complete data cell counts, and draws of the complete data cell counts are made conditional on the cell probabilities and observed cell counts. Estimation under monotonicity restricts complete cell counts and probabilities associated with "harmed" markers and outcomes to be equal to 0; estimation under stochastic monotonicity restricts the cell probabilities associated with the harmed markers and outcomes to be less than the cell probabilities associated with responsive markers and outcomes by rejecting draws that fail to meet this criterion. Further details of the Gibbs sampler are not provided because of space limitations.

3.3 Choosing Between the Missingness Mechanisms

To choose between the latent ignorable and MCAR missingness mechanisms, we can compute the Deviance Information Criterion (DIC) of Spiegelhalter et al. [19]. The DIC measure accounts for the fact that, in a hierarchical framework, the number of effective parameters may be unclear: the random effects associated with each subject may "count" as approximately one parameter if the between-variance estimates are large (small degree of shrinkage), and as nearly zero parameters if the between-variance estimates are small (large degree of shrinkage). Although we do not entertain a Bayesian hierarchical model here, we have a similar issue in that the number of parameters is unclear given that not all are fully identified.

4 Application to the Collaborative Initial Glaucoma Treatment Study

We apply the proposed models to an analysis of the Collaborative Initial Glaucoma Treatment Study (CIGTS) [13]. Glaucoma is an eye disease caused by increased intraocular pressure (IOP) that can result in reduced vision or blindness. The CIGTS was a clinical trial that compared the effects of eye surgery (treatment) against the standard practice of medication (control) to reduce or stop visual field loss. Because visual field loss is caused by increased IOP, one of the major secondary outcomes of interest is reduction in IOP. Here we consider reduction in IOP below 18 mmHg after 96 months of follow-up, based on previous work that has shown IOP of less than 18 mmHg at every time point during at least 6 years of follow-up was associated with a reduced likelihood of visual field loss [1]. Because of the extensive follow-up time, we wish to determine if early reductions in IOP could serve as a marker for much late reductions; hence the surrogate marker was reduction in IOP below 18 mmHg after 12 months of follow-up. However, such an analysis suffers from a substantial amount of missing outcome data due to the long follow-up period. Because the cause of the missingness is due to dropout for unknown reasons, we do not restricting the missingness patterns in the analysis.

The observed data is given in Table 1. Of 574 subjects with IOP measured at 12 months, only 228 had fully observed data (IOP also measured at 96 months). For

Table 1 Collaborative Initial Glaucoma Treatment Study: Observed Data.

		Control				Treatment			
		Reduced IOP at 96 Months				Reduced IOP at 96 Months			
		No	Yes	Missing		No	Yes	Missing	
Reduced IOP at 12 Months	No	28	29	69	145	11	8	35	65
	Yes	14	55	97	147	10	73	144	216
		42	84	166	292	21	81	179	282

fully observed subjects on the control (drug only), 66.7 % had reduced IOP to below 18 mmHg at 96 months; 79.4 % of fully observed subjects on the treatment arm had reduced IOP at 96 months, yielded an estimated causal effect of treatment (CET) of .127 (90 % CI .030,.224). Reduced IOP at 12 months was observed for 58.4 % of subjects who were fully observed on the control arm, versus 54.8 % of subjects who did not have 96-month IOP measures. For subjects on the treatment arm, 80.4 % of fully observed subjects had reduced IOP at 12 months, versus 81.4 % of subjects without 96-month IOP measures.

We fit an MCAR model and a latent ignorable model, as well as a model for the fully observed data, under the monotonicity and non-monotonicity assumption, as well as the "stochastic monotonicity" assumption that only assumes the treatment more likely to be helpful than harmful. Each model is fit using a single chain of 100,000 draws after a burn-in of 1,000. We consider uniform priors of the form $a_{j|i} = 1$ for all i, j and $b_{i+k} = 1$ for all i, k for the latent ignorable model and $a_{ij+} = 1$ and for all i, j and $b_{++k} = 1$ for all k for the MCAR model, and assess sensitivity to the prior by also considering of Jeffreys-type prior of the form $a_{1|1} = \cdots = a_{4|4} = 1/2$ and $b_{1+1} = \cdots = b_{4+4} = 1/2$ and $a_{11+} = \cdots = a_{44+} = 1/2$ and $b_{++1} = \cdots = b_{++4} = 1/2$. Results are given in Table 2. Table 3 provides the DIC measures for the latent ignorable and MCAR models.

Based on DIC, the best fit is provided by the latent ignorable model under stochastic monotonicity; similar fit is provided by the MCAR model under monotonicity and the uniform Dirichlet prior. Particularly poor fit is evidenced by the MCAR model under non-monotonic assumptions, as evidenced by the discrepancy between the CE estimator from the model and the identifiable estimate (.127) obtained from the fully observed data; use of the Jeffreys-type prior improved the fit to some degree. The best-fitting latent ignorable model under stochastic monotonicity and MCAR model under monotonicity had little sensitivity to the prior assumptions and gave broadly similar results. In particular, early reduction of IOP appears to be at best a modestly useful surrogate marker from a causal perspective, with only a weak association between the 8-year causal effect on IOP and the 12-month causal effect on IOP. The point estimate of the associative proportion is somewhat greater under the latent stochastic monotonicity assumption than under the MCAR full monotonicity assumption, although the 90 % credible interval can and does include 0, indicating some evidence of interactions in causal effect within the principal strata [5].

5 Discussion

This manuscript considers a principal stratification approach to assess surrogacy for dichotomous markers and outcomes when missing data is present for the outcome, extending the work of Li et al. [10] to accommodate missingness under latent ignorability assumptions. The principal strata are defined by the joint distribution of the surrogate marker under both treatment and control, with the quality of the

Table 2 Analysis of Collaborative Initial Glaucoma Treatment Study: posterior mean, posterior mode, and 90 % credible intervals for causal treatment effect (CE), associative effect (AE), associative proportion (AP), and common associative proportion (CAP). First row for Dirichlet prior set uniformly to 1; second row for Dirichlet prior set uniformly to 1/2.

	T Fully Observed			MCAR			Latent Ignorability		
	Mean	Mode	CI	Mean	Mode	CI	Mean	Mode	CI
Monotonicity									
CE	.130	.125	.060-.209	.126	.118	.058-.200	.148	.141	.069-.233
	.120	.112	.044-.203	.115	.111	.042-.195	.133	.123	.039-.234
AE	.045	.009	.004-.109	.040	.008	.003-.097	.052	.030	.004-.118
	.042	.003	.000-.121	.035	.002	.000-.102	.050	.003	.000-.142
AP	.339	.192	.034-.731	.316	.119	.030-.698	.348	.322	.039-.726
	.339	.020	.004-.863	.308	.020	.003-.830	.362	.023	.005-.884
CAP	.134	.025	.011-.323	.127	.023	.010-.315	.159	.031	.013-.376
	.131	.009	.001-.383	.117	.008	.001-.354	.164	.013	.001-.492
No Monotonicity									
CE	-.024	-.022	-.119-.070	.015	.018	-.076-.106	.073	.080	-.054-.185
	.032	.030	-.064-.127	.060	.062	-.032-.151	.061	.085	-.121-.205
AE	-.027	-.018	-.130-.074	.025	.019	-.071-.121	.068	.084	-.066-.187
	.031	.036	-.080-.142	.066	.071	-.044-.171	.062	.094	-.109-.211
AP	.827	.915	-3.482-5.125	.517	.972	-4.379-5.793	.853	.849	-2.189-3.850
	.845	.979	-3.734-5.427	.987	.978	-2.446-4.430	.795	.901	-2.341-3.866
Stochastic Monotonicity									
CE	.023	.028	-.060-.107	.044	.048	-.037-.125	.116	.120	.031-.205
	.056	.061	-.034-.144	.079	.074	-.003-.163	.125	.124	.027-.232
AE	-.015	-.026	-.100-.072	.006	.002	-.075-.086	.055	.047	-.025-.141
	.010	.014	-.084-.103	.034	.029	-.056-.124	.057	.034	-.031-.156
AP	.650	.578	-4.060-5.902	.433	.690	-2.863-3.900	.426	.551	-.425-.884
	.400	.773	-2.744-3.486	.367	.791	-1.538-1.331	.409	.552	-.483-.912

Table 3 DIC measures for various models accounting for missing data. First row for Dirichlet prior set uniformly to 1; second row for Dirichlet prior set uniformly to 1/2.

	MCAR	Latent Ignorability
Monotonicity	63.68	67.64
	64.13	67.97
No Monotonicity	79.06	66.37
	69.19	66.62
Stochastic Monotonicity	77.53	63.66
	68.48	63.25

surrogate being the CET that is associated with strata in which the surrogate marker is impacted by treatment. Latent ignorability assumes that the conditional distribution of the potential outcomes within the principal strata are independent of the outcome missingness but allows for the possibility that the marginal distribution of the surrogate marker or treatment outcome is associated with outcome missingness. This is a weaker assumption than MCAR in which missingness is independent

of both the surrogate and the true outcome, and is identifiable at the "complete data" (counterfactual) level, in contrast to a fully non-ignorable model which would require postulating non-identified parameters for the unobserved outcomes at this complete-data level. A unique aspect of our approach (to our knowledge) is that we allow for the possibility that missingness has a counterfactual component, one that might differ between the treatment and control due to differential dropout.

A variety of extensions to this work can be considered. Our focus is on assessing and ameliorating the effect of missingness on inference about binary surrogate measures and outcomes in a counterfactual setting, and in the process we have focused on a relatively simple Dirichlet prior formulation for the cell parameters. The work of Li, Taylor, and Elliott [10] and Li et al. [11] considered a log-linear parameterization that was capable of incorporating a priori assumptions about positive correlations between the surrogate marker and final outcome in a more refined fashion than the model considered here, particularly when the monotonicity assumption is relaxed. Extensions to continuous surrogate measures and outcomes are also possible and are the focus of current work.

References

[1] AGIS Investigators.: The Advanced Glaucoma Intervention Study (AGIS) 7: The relationship between control of intraocular pressure and visual field deterioration. American Journal of Ophthalmology. **130**, 429–440 (2000)

[2] Burzykowski, T., Molenberghs, G. and Buyse, M.: The Evaluation of Surrogate Endpoints. Springer-Verlag, New York (2005)

[3] Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H.: The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics **1**, 49–67 (2000)

[4] Chen, T.T., Simon, R.M., Korn, E.L., Anderson, S.J., Lindblad, A.S., Wieand, H.S., Douglass, H.O. Jr, Fisher, B., Hamilton, J.M. and Friedman, M.A.: Investigation of disease-free survival as a surrogate endpoint for survival in cancer clinical trials. Communications in Statistics: Theory and Methods **27**, 1363–1378 (1998)

[5] Elliott, M.R., Raghunathan, T.E. and Li, Y.: Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. Biostatistics **11**, 353–372 (2010)

[6] Frangakis C. and Rubin D.B.: Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. Biometrika **86**, 365–379, 1999

[7] Frangakis, C. and Rubin, D.B.: Principal stratification in causal inference. Biometrics **58**, 21–29 (2002)

[8] Gilbert, P.B. and Hudgens, M.G.: Evaluating candidate principal surrogate endpoints. Biometrics **64**, 1146–1154 (2008)

[9] Gustafson, P.: Bayesian inference for partially identified models. The International Journal of Biostatistics. **6**:17 (2010)

[10] Li, Y., Taylor, J.M.G. and Elliott, M.R.: A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. Biometrics **66**, 523–531 (2010)

[11] Li, Y., Taylor, J.M.G., Elliott, M.R. and Sargent D.J.: Causal assessment of surrogacy in a meta-analysis of colorectal cancer trials. Biostatistics **12**, 478–492 (2011)

[12] Lin, D.Y., Fischl, M.A. and Schoenfeld, D.A.: Evaluating the role of CD4-lymphocyte counts as surrogate endpoints in Human Immunodeficiency Virus clinical trials. Statistics in Medicine **12**, 835–842 (1993)
[13] Musch D.C., Lichter P.R., Guire K.E., Standardi C.L. and CIGTS Investigators.: The Collaborative Initial Glaucoma Treatment Study (CIGTS): Study design, methods, and baseline characteristics of enrolled patients. Ophthalmology **106**, 653–662. (1999)
[14] Peng, Y., Little, R.J.A. and Raghunathan, T.E.: An extended general location model for causal inferences from data subject to noncompliance and missing values. Biometrics **60**, 598–607 (2004)
[15] Prentice R.L.: Surrogate endpoints in clinical trials: Definition and operational criteria. Statistics in Medicine **8**, 431–440 (1989)
[16] Robins, J.M. and Greenland, S.: Identifiability and exchangeability for direct and indirect effects. Epidemiology **3**, 143–155 (1992)
[17] Rosenbaum, P.R.: The consequences of adjustment for a concomitant variable that has been affected by the treatment. Journal of the Royal Statistical Society **A147**, 656–666 (1984)
[18] Rubin D.B.: Formal modes of statistical inference for causal effects. Journal of Statistical Planning and Inference **25**, 279–292 (1990)
[19] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van Der Linde, A.: Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society **B64**, 583–639 (2002)
[20] Taylor, J.M.G., Wang, Y. and Thiébaut, R.: Counterfactual links to the proportion of treatment effect explained by a surrogate marker. Biometrics **61**, 1102–1111 (2005)
[21] Taylor, L. and Zhou, X. H.: Multiple imputation methods for treatment noncompliance and nonresponse in randomized clinical trials. Biometrics **65**, 88–95 (2009)

Assessment of Treatment-Mediated Correlation Between a Clinical Endpoint and a Biomarker

Peter H. Hu

Abstract There is increasing need to identify biomarkers (BMKs) responding early to drug treatment to help decision making during clinical development. One of the statistical metrics often involved in screening such BMKs from a single study is the assessment of correlation between a candidate BMK and a primary clinical endpoint. In this chapter, some drawbacks in relying on simple regression models for such an investigation will be criticized first, followed by a real example to demonstrate the danger of relying on static data to assess such a correlation. A theoretical justification will then be given to promote the idea of pursuing treatment-mediated correlation patterns. The rest of this paper will then be focused on how to estimate correlation under this preferred metric from data with parallel-group design and time-to-event (T2E) being the primary clinical endpoint. A jointly modeling framework of T2E and longitudinally measured BMK will then be introduced, with explanation in details how to parameterize the joint model and interpret some key parameters. By comparing the performances of three different models, the results from the analysis of an AIDS trial will be presented to demonstrate the benefit of joint modeling of T2E and BMK, followed by some brief discussions.

1 Introduction

Of various criteria to evaluate biomarker (BMK) use for clinical development, the correlation, or in general speaking, association, between a candidate BMK and a primary clinical endpoint of interest, is the most popular statistical metric to be assessed. Unlike cross-sectional data where snapshot correlation is the only pattern that can be pursued from the data, clinical trial data are dynamic and many

P.H. Hu (✉)
ED-GBS, Bristol-Myers Squibb, Princeton, NJ, USA
e-mail: Peter.Hu@bms.com

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_9,

different correlation metrics are assessable. However, often only treatment-mediated correlation pattern between a BMK and primary clinical endpoint is interpretable and meaningful in terms of how treatment effect observed on a BMK can be used to gauge treatment effect on the primary clinical endpoint that is time consuming to observe. See a good example from a meta-analysis evaluating statin treatment [2]. Although meta-regression is a putative approach to disclose how drug treatment effect on a BMK of interest corresponds to the treatment effect on the primary clinical endpoint, it is unrealistic to have abundant resources of data to conduct meta-analysis for such an investigation at early clinical space. Rather, such an objective has often to be sought from a single, small study. In Sect. 2, the challenges of exploring such a relationship from a single study will be reviewed and some drawbacks of relying on simple linear regression to explore the correlation between a primary clinical endpoint and a longitudinally measured BMK will be criticized, followed by an example in Sect. 2.1 to demonstrate the danger of relying on static data to assess such a correlation. How the information about the intrinsic systematic correlation pattern in the whole data is utilized in various correlation metrics will then be disclosed theoretically in Sect. 2.2. In Sect. 3, a framework of joint modeling of a primary clinical endpoint and a longitudinally measured BMK will then be introduced. Parameterization of the joint model with time-to-event (T2E) being the primary clinical endpoint is to be explained in detail in Sect. 3.1, as well as interpretation of the key parameters. An AIDS clinical trial data will be analyzed in Sect. 3.2 using three different models. Their performances will be compared to demonstrate the benefit of jointly modeling T2E and BMK, followed by some brief discussions in Sect. 4.

2 Abuse of Correlation Metric in Clinical Trial Arena

In the case of meta-analysis, treatment effects on BMK and the primary clinical endpoint can be estimated from each individual studies as paired data and the correlation pattern between the two treatment effects can be visualized across studies [2]. In a single study, however, such a correlation pattern is not visualizable from data with parallel design. As an ad hoc approach to get around this disadvantage, crossover design can be considered to evaluate the correlation between the two types of treatment effects at subject level since each subject can receive two or more different treatments. Due to the difficulty in visualizing treatment-mediated data for correlation assessment from a single study, two naïve approaches are often used—evaluating correlation between a BMK and the primary clinical endpoint for each time point or for each treatment group. There are some drawbacks in these naïve approaches. Even the study design allows a BMK and the primary clinical endpoint to be measured with synchronization, the pattern of correlation between them at each time point, as a stochastic snapshot, can be dramatically different from treatment-mediated systematic pattern, such as change

from baseline. Similarly, correction pattern within each treatment group is not necessarily treatment-mediated. These issues still exist even when all time points and treatment group data are pooled together for correlation assessment if no meaningful contrast is applied to convert the flat data into treatment-mediated dynamic data, i.e., change from baseline or difference from placebo or both. An example is given in Sect. 2.1 to show the danger of relying on static data to assess correlation between a BMK and the primary clinical endpoint.

2.1 The Danger of Relying on Static Data to Assess Correlation

A 4-treatment sequence$\times$3-period crossover study was conducted to evaluate renal blood flow (RBF) among 12 patients. Each patient received 3 out of 4 treatments, placebo, drug A with two doses, A_1 and A_2, and generic drug L. RBF was measured on two assay platforms, G and E_1. Assay platform G is considered the conventional gold standard but it is invasive and requires time series collection of multiple urine samples in order to calculate RBF. E_1 is an image-based new assay that is noninvasive and only one time measurement is needed. Due to the infeasibility of assay platform G, only one post-dose measure of RBF was conducted on both platforms G and E_1. Here RBF data from assay platform G is considered as the primary clinical endpoint and the new assay platform E_1 as the BMK.

The study results indicated that all active drug treatments, A_1, A_2, and L, significantly boosted RBF compared to placebo on both assay platforms (data not shown). Consequently it was of interest to understand whether the same biology was measured on both assay platforms by exploring how the RBF data are correlated between the two assay platforms. Two approaches were proposed for this investigation but, surprisingly, the correlation patterns are dramatically different.

Shown in Fig. 1 are the treatment-specific, as well as all treatment-pooled, correlation patterns between the two assay platforms with static data, that is, the one-time readout of RBF. It's surprising that a negative trend of correlation is observed, with statistical significance achieved in both doses of drug A treatments. Given the fact that all active drug treatments boosted RBF in both assay platforms compared to placebo and it is believed the same biology was measured in both assay platforms, it is very difficult to interpret this negative correlation. Increase in sample size by pooling all treatment data together does not seem to help at all.

Taking a glimpse of the correlation pattern in treatment group A_1, it looks apparently subject # 9 is an outlier. Of course this subject contributed to the negative correlation due to the highest RBF readout in assay platform G and lowest in assay platform E_1. But this negative correlation pattern still remains when this data point is removed. In fact, this subject showed highest RBF on assay platform G as well when receiving placebo. And when taking placebo, this subject's RBF readout is the 3rd lowest on assay platform E_1, suggesting that, if the correlation metric is constructed using placebo adjusted RBF, then the behavior of this data point may be different.

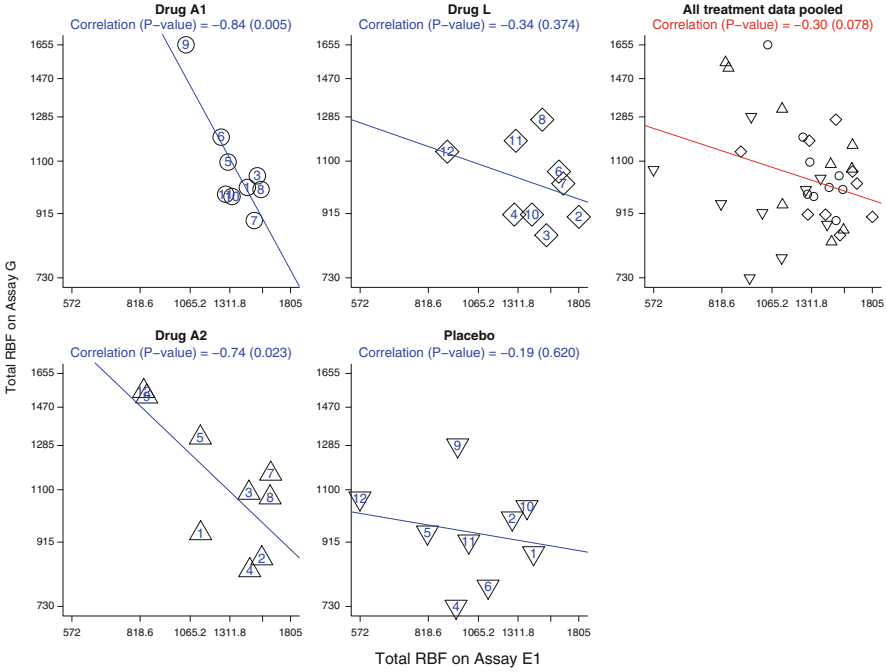


Fig. 1 The x-axis and y-axis are the one-time RBF readout on assay platforms $E_1$and G, respectively. Each type of symbol represents a treatment group and the numbers printed within each symbol are the patient IDs. Correlation patterns between the two assay platforms are plotted separately for each treatment group, as well as pooled together. Printed on top are the estimates of Pearson correlation coefficient and the associated p-values. The blue line corresponds to the slope of the best linear fit of the paired data.

The endpoint applied to construct the correlation metric shown in Fig. 2 is the ratio of RBF readout for an active treatment over placebo and A_2 vs A_1 (for assessment of dose effect of drug A). It can be seen clearly that this correlation pattern, using change from placebo as dynamic data, is dramatically different from the negative trend in Fig. 1. Unfortunately, due to the limitation of study design and very small sample size, no statistical significance at level of 0.1 was reached for this positive trend of the correlation. But it is obvious the patterns shown in these two figures are completely different from each other.

The above example implies not all correlation metrics are consistent with each other and helpful for drug development. To dissect why the two correlation metrics showed dramatically different behaviors, theoretical explorations, from the point of view how information about the intrinsic systematic correlation pattern in the whole data is utilized in the two correlation metrics, were investigated further.

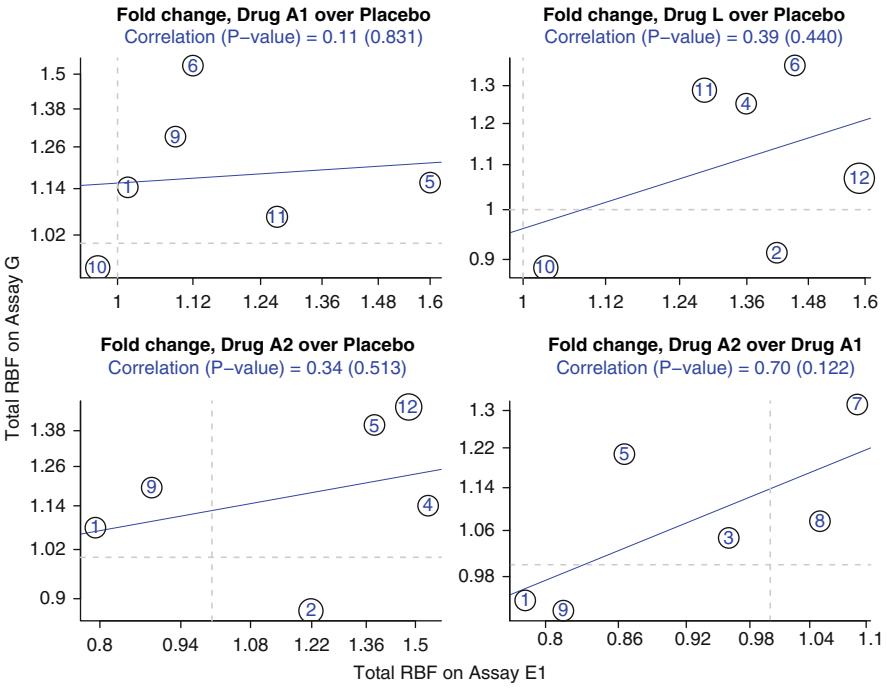


Fig. 2 The x-axis and y-axis are the fold changes of RBF readout for an active drug treatment over placebo on assay platforms E_1 and G, respectively. Correlation patterns between the two assay platforms are plotted separately for each active treatment. Other settings are the same as Fig. 1.

2.2 Information Usage in Various Correlation Metrics

Sample correlation between two data matrices $\mathbf{U}$ and $\mathbf{V}$ can be interpreted geometrically in a linear space as $\cos(\theta)$, where θ is the angle between them [4]. The goal of this theoretical investigation is to express $\mathbf{U}$ and $\mathbf{V}$ in various correlation metrics as a function of the entire data $\mathbf{X}$ to disclose how the intrinsic correlation information imbedded on $\mathbf{X}$ is represented in $\cos(\theta)$. Mimic the data structure of the above example, let each of n subjects receive both active treatment p and control treatment q under both assay platforms u and v, then the entire outcome data can be expressed as $\ln(\mathbf{X}_{n\times 4}) = (\mathbf{u}_p | \mathbf{u}_q | \mathbf{v}_p | \mathbf{v}_q)$. Note that, if $\ln(\mathbf{X})$ is centered at its mean vector, then $\cos(\theta)$ will be exactly the Pearson correlation coefficient estimate as calculated according to (1) and $\ln(\mathbf{X})^T \ln(\mathbf{X})$ is proportional to the sample variance–covariance matrix $\widehat{\text{cov}[\ln(\mathbf{X})]}$ denoted in (1)

Table 1 Data input for 3 correlation metrics and the estimated correlations

Data & Correlation	Pooled Treatment	Active Treatment	Treatment Effect
In assay **U**	$\mathbf{ln}(\mathbf{X})\begin{pmatrix}1&0&0&0\\0&1&0&0\end{pmatrix}^T$	$\mathbf{ln}(\mathbf{X})\begin{pmatrix}1&0&0&0\end{pmatrix}^T$	$\mathbf{ln}(\mathbf{X})\begin{pmatrix}1&-1&0&0\end{pmatrix}^T$
In assay **V**	$\mathbf{ln}(\mathbf{X})\begin{pmatrix}0&0&1&0\\0&0&0&1\end{pmatrix}^T$	$\mathbf{ln}(\mathbf{X})\begin{pmatrix}0&0&1&0\end{pmatrix}^T$	$\mathbf{ln}(\mathbf{X})\begin{pmatrix}0&0&1&-1\end{pmatrix}^T$
$\cos(\theta)$	$\frac{h+j}{\sqrt{a+b}\sqrt{c+d}}$	$\frac{j}{\sqrt{a}\sqrt{c}}$	$\frac{h+j-(g+i)}{\sqrt{a+b-2e}\sqrt{c+d-2f}}$

$$\cos(\theta)=\frac{\mathrm{tr}\left(\mathbf{U}^T\mathbf{V}\right)}{\sqrt{\mathrm{tr}\left(\mathbf{U}^T\mathbf{U}\right)}\sqrt{\mathrm{tr}\left(\mathbf{V}^T\mathbf{V}\right)}},\widehat{\mathrm{cov}[\mathbf{ln}(\mathbf{X})]}=\begin{array}{c|cccc} & \underline{\mathbf{u}_p} & \underline{\mathbf{u}_q} & \underline{\mathbf{v}_p} & \underline{\mathbf{v}_q} \\ \mathbf{u}_p & a & e & j & i \\ \mathbf{u}_q & e & b & g & h \\ \mathbf{v}_p & j & g & c & f \\ \mathbf{v}_q & i & h & f & d \end{array} \tag{1}$$

Summarized in Table 1 are the three correlation metrics proposed, their input endpoint data expressed as functions of data **X** and the corresponding correlations estimates.

As it can be seen from the expressions of $\cos(\theta)$, the correlation information represented as the ten unique elements, a through j, in $\widehat{\mathrm{cov}[\mathbf{ln}(\mathbf{X})]}$ denoted in (1) is only fully utilized in the estimate for correlation in treatment effects. If correlation is sought only from active treatment data, then any correlation information related to placebo will be ignored, leading to substantial loss of information. From this point of view, the negative correlation detected from the treatment group with drug *A* in the above example may be biased. More surprisingly, intuitively there appears to have no loss of information if pooled data are applied for correlation estimation. But, as it can be seen from the expression of $\cos(\theta)$, only six out of the ten unique elements in the sample variance–covariance matrix is utilized, suggesting meaningful contrasts should be applied to convert pooled data into dynamic data to avoid loss of information. Simply pooling all treatment data together cannot guarantee to have all correlation information retained.

Although correlation in subject-level treatment effects can be visualized and realized in crossover studies as treatment-mediated correlation measure, this ad hoc approach does not work for parallel group comparisons because no individual patient level treatment effect is observable. For treatment-mediated associations between a primary clinical endpoint and a longitudinal BMK with parallel group design or crossover design with repeated measures, a joint modeling between the clinical endpoint and BMK has to be applied.

3 Joint Modeling of a Clinical Endpoint and a Longitudinally Measured BMK

Joint modeling of two correlated endpoints, X and Y, can be realized following the theory in latent variable model. In general, two assumptions are made. First, it assumes there exists a latent variable or random effect w that underlies the process of both X and Y. Second, it assumes X and Y are independent of each other conditioning on w. Provided $X|w$ and $Y|w$ can be modeled parametrically from the data with likelihood $l(X|w)$ and $l(Y|w)$, respectively, the joint likelihood of X and Y, $l(X,Y)$, can then be computed by integrating $l(X,Y|w) = l(X|w)\,l(Y|w)$ over the distribution of w as nuisance parameters. To obtain the inference about the association between treatment effect on BMK and the primary clinical endpoint, an extra step called bridging is imposed, that is, the subject level mean profile predicted from the model for $X|w$ is included as one of the covariates during modeling $Y|w$. Considering X as a normally distributed, longitudinally measured BMK and Y as T2E, parameterization of the joint likelihood is illustrated below.

3.1 Construction of Joint Likelihood of a Longitudinally Measured BMK and T2E

To parameterize the joint likelihood of a longitudinally measured BMK and T2E, we observe

1. Longitudinally measured BMK outcome on subject i at time t_{ij} as $x_i(t) = \{x_i(t_{ij}),\ j = 1,\ldots k_i\}$. Denote $\mathbf{z}_i$ a p-dimensional vector of baseline covariates, including treatment indicator, for subject $i = 1,\ldots,n$, then the mean profile of $x_i(t)$ can be modeled as $x_i(t) = m_i(t) + \varepsilon_i(t), \varepsilon_i(t) \sim \mathcal{N}\left(0,\sigma^2\right)$, where

$$m_i(t) = \gamma_{0i} + \gamma_{1i}t + \gamma_{2i}t^2 + \cdots \gamma_{di}t^d + \mathbf{z}_i^T\eta \tag{2}$$

   where $\gamma_i = (\gamma_{0i},\gamma_{1i},\ldots,\gamma_{di})^T \sim \mathcal{N}_d\left(\mu_\gamma,\Sigma_\gamma\right)$ is a d-dimensional random effect with PDF denoted as $g(\cdot)$. Note that the random effects presented in (2) as high-order polynomials are to provide better fitting in case some subjects show highly nonlinear longitudinal trajectories, as recommended in [5]. The likelihood for BMK process can then be written as $f\left(x_i(t_{ij})\,\middle|\,\gamma_i;\mathbf{z}_i,\eta,\sigma^2\right)$.
2. T2E data $y_i = (T_i,\delta_i)$ for subject i, where $T_i = \min(T_i^*,C_i)$, with T_i^* and C_i being the true event time and censoring time, respectively. δ_i is the event indicator, i.e., $\delta_i = I(T_i^* \le C_i)$. Denote $\mathbf{w}_i$ a q-dimensional vector of baseline covariates, including treatment indicator, for subject $i = 1,\ldots,n$. Denote further $\mathcal{M}_i(t) = \{m_i(\lambda), 0 \le \lambda < t\}$ the history of the true unobserved longitudinal BMK process

up to time point t, $h_0(t)$ the baseline risk function. The risk function at time t can then be modeled as

$$h(t\,|\mathscr{M}_i(t);\mathbf{w}_i) = h_0(t)\exp\left[\mathbf{w}_i^T\,\alpha + \beta m_i(t)\right] \tag{3}$$

The likelihood for the T2E process is then

$$u_i(T_i,\delta_i\,|\gamma_i;\mathbf{w}_i,\alpha,\beta,\eta) = \left[h(T_i\,|\mathscr{M}_i(T_i);\mathbf{w}_i,\beta,\alpha)\right]^{\delta_i} S_i(T_i\,|\mathscr{M}_i(T_i);\mathbf{w}_i,\beta,\alpha)$$

where $S_i(T_i\,|\mathscr{M}_i(T_i);\mathbf{w}_i,\beta,\alpha) = \exp\left\{-\int_0^{T_i} h_0(t)\exp\left[\mathbf{w}_i^T\alpha + \beta m_i(t)\right]dt\right\}$.

The complete joint likelihood for subject i is then

$$\int_{-\infty}^{\infty}\left[\prod_{j=1}^{k_i} f\left(x_i(t_{ij})\,|\gamma_i;\mathbf{z}_i,\eta,\sigma^2\right)\right] u_i(T_i,\delta_i\,|\gamma_i;\mathbf{w}_i,\alpha,\beta,\eta)\,g\left(\gamma_i;\mu_\gamma,\Sigma_\gamma\right)d\gamma_i \tag{4}$$

Treatment-mediated association between BMK and T2E, as well as treatment effects on BMK and T2E, respectively, can then be obtained by maximizing (4). Although theoretically maximization of (4) can be realized using EM algorithm [7], a hybrid of EM and quasi-Newton algorithms, implemented recently as an R contributed package called JM [6], is much more efficient. Some handy utility functionalities there enable more choices of T2E models and numeric integration methods and allow estimating expected survival probability and residuals via resampling.

Revisit the two core sub-models, (2) and (3), in the joint modeling process, it can be seen that some key inferences can be obtained in this joint modeling framework from three parameters, η (treatment effect on BMK), α (treatment effect on T2E), and β (association strength between T2E and treatment effect on BMK), such that $\beta\eta + \alpha$ is the combined treatment effect on T2E. Some further discussion about these effects will be provided in Sect. 4. Below an anti-AIDS clinical trial data will be applied to fit three models and the benefits of using this joint model are illustrated.

3.2 An Example

A randomized, parallel-group clinical trial among 467 HIV infected patients who did not respond to AZT in the previous trial participated in this study. The aim of this study was to compare the efficacy and safety of two alternative anti-retroviral drugs (variable name: drug), ddI (Zalcitabine), *and* ddC (Didanosine). During the study follow-up, 188 death events (variable name: event) occurred. $CD4^+$ cell counts (variable name: CD4) scheduled at baseline (Month 0), 2, 6, 12, and 18 months (variable name: obstime) post-randomization, as well as

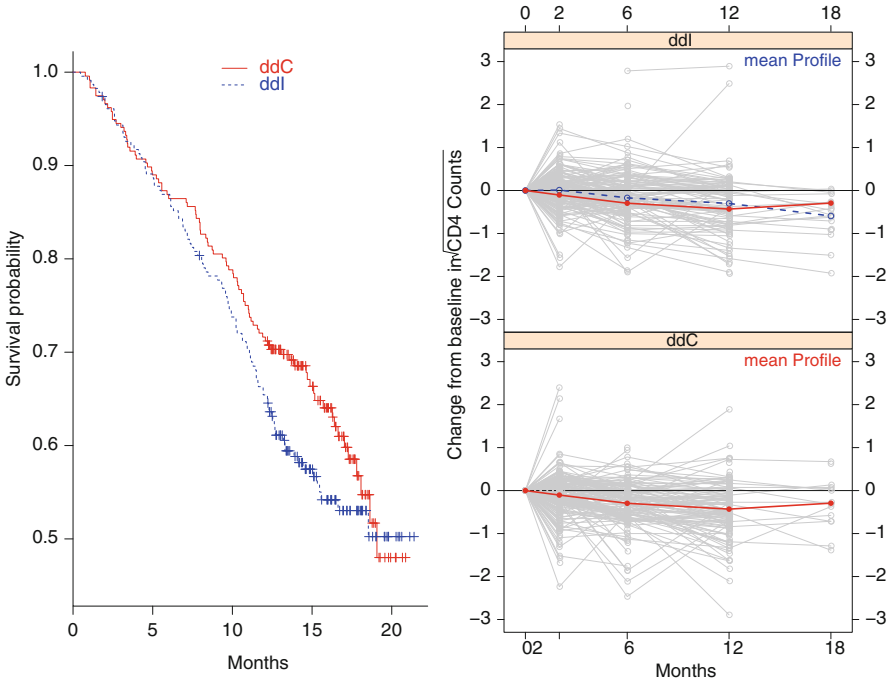


Fig. 3 Plot at the left panel is the Kaplan-Meier survival curve of ddI vs ddC. Plot at the right panel is the individual profile of change in $\sqrt{\text{CD4}^+\text{cell counts}}$ from baseline (gray lines). Mean profiles of such a change are presented by colorful lines.

time-to-death (variable name: Time), were recorded. Additional three important baseline explanatory binary variables for AIDS diagnosis at study entry (yes vs no, variable name: prevOI), AZT stratum (failure vs intolerance, variable name: AZT) and gender (male vs female, variable name: gender) are also included in the analysis. Details about this data can be found from [3].

The first glimpse of treatment-mediated association between time-to-death and CD4 cell counts can be taken from Fig. 3. Lower probability of survival with ddI treatment compared with ddC is observed from the left panel. From the right panel, $CD4^+$ cell counts decreased over time. When compared the mean profile curves over time between the two treatment groups, it can be seen that the treatment group with ddI showed slower pace of decrease of $CD4^+$ cell counts. This pattern suggests the correlation in treatment effects (ddI vs ddC) on time-to-death and $CD4^+$ cell counts should be negative.

To assess treatment-mediated association strength between time-to-death and $CD4^+$ cell counts, the following three models are considered to compare with each other. All models contain the three explanatory variables, prevOI, AZT, and gender, as baseline covariates.

Table 2 Comparisons in point estimate (SE) among 3 models[@]

Parameter	Point Estimate (SE)		
	Naïve Model	2-stage Model	Joint Model
Treatment effect on T2E ($\hat{\alpha}$)	0.222 (0.147)	0.233 (0.148)	0.274 (0.144)
Treatment effect on BMK ($\hat{\eta}$)		0.040 (0.024)	0.041 (0.024)
Association between T2E and BMK ($\hat{\beta}$)	−0.045 (0.137)	−0.273 (0.339)	−0.400 (0.150)

The treatment effect (SE) on T2E estimated from a Cox model with treatment being the only covariate is 0.217 (0.146)

1. *Joint model:* The LDA sub-model and Cox sub-model are specified as follows:
   LDA sub-model: Outcome variable derived as ΔCD4 = change in square root of CD4$^+$ cell counts from baseline, with fixed effects including additional covariates—drug, obstime and drug × obstime, random effect including only obstime;
   Cox sub-model: With additional covariates for drug and the subject-level mean ΔCD4$^+$ predicted from the above LDA sub-model at event or censoring time T_i for subject $i = 1, \ldots, n$.
2. *2-Stage model:* The same as the joint model setup described above but the two sub-models are estimated separately except jointly;
3. *Naïve model:* Same as the above Cox sub-model except the predicted subject-level mean $\Delta\text{CD4}(T_i)$ replaced with raw ΔCD4 and estimated using counting process.

By using the R package JM, the key parameter estimates are displayed in Table 2. It can be seen that the joint model is consistent with the other two models in terms of the treatment effects on time-to-death and CD4$^+$ cell counts, respectively, from both the magnitude and direction of the two effects. The direction of the trend of the treatment-mediated association between time-to-death and CD4$^+$ cell counts is consistent among the three models, which is aligned with what is implied from Fig. 2. However, there are striking differences between the joint model and the other two ad hoc approaches in terms of the magnitude of the association, as well as the estimation efficiency. The joint model is more powerful than either of the ad hoc methods in these two aspects. See Sect. 4 about some further discussions.

4 Discussion

As long as there is only one event or censoring for each subject during the study follow-up, it is in general not a problem to ignore measurement error brought in due to repeated measure of a BMK as a time-varying covariate in counting process-based estimation of Cox-model parameters. This is because only the last time interval, from the last time the measurement of BMK was made to the time of event or censoring, contributes to the partial likelihood, based on which parameter

estimations are made. The underlying assumption is all information about the BMK is absorbed in the last measurement, which, if not valid, may lead to substantial loss of information. This may explain why the estimated correlation strength is a lot weaker in the naïve model. Furthermore, BMK data are more likely to be missing at later scheduled time points, such that the longer the survival time, the less reliable the predicted mean profile due to missing data. Worst of all, mechanism of missing at random is unlikely to be valid under such a scenario. These impacts may explain why the correlation strength estimated from the naïve model or 2-stage model seems attenuated compared to the joint model. In addition, Between-subject variability of BMK is not taken into consideration at the 2nd step of modeling for survival time, resulting in attenuated and/or distorted assessment of correlation in the 2-stage modeling. The joint model, on the other hand, is not only unaffected by any of these impacts, but also may provide a plausible solution to LDA with informative missing data, with inferences for BMK response made directly and no need to impute missing data explicitly.

The combined total treatment effect on T2E inferred from the two core sub-models, (2) and (3), is $\beta\eta + \alpha$. If the magnitude of $\beta\eta + \alpha$ is greater than α, then a trial design by considering simultaneously information about BMK and T2E could be more cost-effective than based on T2E alone. See [1] for details. Also note that β, estimated from a single study, does not have exactly the same interpretation as correlation between treatment effect on BMK and treatment effect on T2E as from meta-analysis [2] although both of them are treatment-mediated correlations. From β it can tell how much effect on T2E is attributed to the treatment effect on BMK, in the sense that, if $\beta\eta$ takes the majority in the combined effect on T2E, then the treatment effect at the end of the trial to be gauged during early interim analysis can be at least $\beta\eta$. If η, the treatment effect on BMK, improves over time, then improvement of the treatment effect at the end of the trial is expected as well. To obtain the correlation estimate from a single study with meaning equivalent to the one from meta-analysis, we may stratify the total study population and perform meta-analysis type of estimation by pretending each stratum as an independent study, which requires more careful plan during the trial design stage than the routine case, i.e., to find an appropriate stratifier and perform stratified randomization at larger sample size.

Acknowledgments I would like to thank the three editors' careful and valuable reviews on the draft of this manuscript and express my appreciations to the Publisher's efforts during the finalization of this manuscript.

References

[1] Chen, L.M., Ibrahim, J.G., Chu, H. (2011). Sample size and power determination in joint modeling of longitudinal and survival data. *Statist. Med. 30, 2295-2309.*

[2] Cholesterol Treatment Triallists' Collaborators (2005). Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90056 participants in 14 randomised trials of statins. *Lancet. 366, 1267-78.*

[3] Guo, X., Carlin, B. (2004). Separate and joint modeling of longitudinal and event time data using standard computing packages. *American Statisticians. 58(1), 1-9.*
[4] Harville, D.A. (1997). Matrix algebra from a statistician's perspective. Springer, New York.
[5] Ibrahim, J.G., Chu, H., Chen, L.M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *J. Clin. Oncol. 28(16), 2796-2801.*
[6] Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *J. Stat. Software. 35(9), 1-33.*
[7] Wulfsohn, M.S., Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics 53, 330-339.*

Biomarker Selection in Medical Diagnosis

Man-Jen Hsu, Yuan-Chin Ivan Chang, and Huey-Miin Hsueh

Abstract A biomarker is usually used as a diagnostic or assessment tool in medical research. Finding a single ideal biomarker of a high level of both sensitivity and specificity is not an easy task; especially when a high specificity is required for a population screening tool. Combining multiple biomarkers is a promising alternative and can provide a better overall performance than the use of a single biomarker. It is known that the area under the receiver operating characteristic (ROC) curve is most popular for evaluation of a diagnostic tool. In this study, we consider the criterion of the partial area under the ROC curve (pAUC) for the purpose of population screening. Under the binormality assumption, we obtain the optimal linear combination of biomarkers in the sense of maximizing the pAUC with a pre-specified specificity level. Furthermore, statistical testing procedures based on the optimal linear combination are developed to assess the discriminatory power of a biomarker set and an individual biomarker, respectively. Stepwise biomarker selections, by embedding the proposed tests, are introduced to identify those biomarkers of statistical significance among a biomarker set. Rather than for an exploratory study, our methods, providing computationally intensive statistical evidence, are more appropriate for a confirmatory analysis, where the data has been adequately filtered. The applicability of the proposed methods are shown via several real data sets with a moderate number of biomarkers.

M.-J. Hsu • H.-M. Hsueh (✉)
Department of Statistics, National ChengChi University, Taipei 11605, Taiwan
e-mail: 95354503@nccu.edu.tw; hsueh@nccu.edu.tw

Y.-C.I. Chang
Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan

Department of Statistics, National ChengChi University, Taipei 11605, Taiwan
e-mail: ycchang@sinica.edu.tw

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_10,

1 Introduction

A biomarker is a biological indicator in showing absence, presence, or the condition of a disease, and it can be used to determine the status of a subject, the effectiveness of a treatment, and so on. Ideally, a biomarker with both high sensitivity and specificity for accurate prediction is expected. However, it is not easy to find such a biomarker in practice. Combining biomarkers provides an alternative to improve the performance of currently available ones. For example, the serum prostate-specific antigen (PSA) is a well-accepted prognostic biomarker to screen for prostate cancer. However, this test has a low specificity and therefore might lead to over-diagnosis. Besides PSA, several potentials are investigated, please see [11]. Nevertheless, no single biomarker among them outperforms PSA, and therefore, more investigators propose the use of a combination of PSA and others. Please see [1, 9].

The receiver operating characteristic (ROC) curve is the most popular graphical tool in evaluation of the diagnostic power of a biomarker. It provides an exhaustive look at the relationship between sensitivity and specificity of a biomarker. The area under the ROC curve (AUC) is proposed for an efficient summarization. In some applications, investigators place their emphasis only on a part of the curve. For example, a high level of specificity is required for a biomarker serving as a population screening tool. As a consequence, a biomarker is assessed on the partial area under the ROC curve (pAUC) in a region of specificity above some level. See [13] and the reviews by [4, 16].

This study focuses on combining multiple continuous-scaled biomarkers into one single diagnostic or predictive rule for disease with emphases on assessment of each biomarker. For better interpretability, we propose the use of a linear combination for information summarization. The discriminatory power of a linear combination of biomarkers is evaluated on the pAUC. The optimal linear combination, which provides the best discriminatory power among all combinations, is the target solution of research interest. In addition to the global predictability, some insights into the importance of an individual biomarker can be obtained from the coefficients. However, it needs to incorporate sampling variation for statistical significance.

In presence of multiple biomarkers, a traditional way is fitting a multiple logistic regression model to the data set for medical diagnosis. For example, see [20]. Alternatively, seeking the maximal discriminatory power, Su and Liu [17] derived the explicit form of the best linear combination in terms of AUC under a binormal model. Following their study, [6] found a solution, that dominates any others in some scenarios. Nevertheless, the dominating scenarios are not universal. Pepe and Thompson [12] and Pepe et al. [14] proposed the use of empirical AUC estimates in finding the optimal linear combination. In our earlier study, we found that not only the analytical derivation but also the computation became much more complicated under the pAUC criterion, please refer [2].

Recently, due to newer and better biotechnology, big data are generated easily and related analytical tools are demanding. In developing a binary classification, which is parallel to a diagnostic rule, several algorithm-based approaches have been

proposed by directly using either AUC or pAUC as the objective function, such as [3, 5, 7, 8, 10, 15, 22, 23]. However, these algorithm-based methods are unable to accommodate statistical evidence into variable selection. It motivates our study in developing some stepwise approaches, embedding adequate statistical tests, to identify important biomarkers for data sets of a moderate size. In which, a biomarker is discarded or selected based on the statistical evidence from data, not only on a computational prospect.

The paper is organized as follows. In Sect. 2, the sample version of the optimal linear combination will be defined. In Sect. 3, some testing procedures for the global and individual discriminatory power will be proposed. Furthermore, two biomarker selection approaches adopting the proposed tests will be developed. Real example analyses are given in Sect. 4. We then conclude this paper with a discussion in Sect. 5.

2 Strong Consistency of the Linear Combination Estimator Maximizing the pAUC

Let $\mathbf{X}$ be a random vector of p biomarkers related to the disease of a subject, and D be the binary disease status, where $D = 1$ indicates a subject from the diseased population, $D = 0$ indicates a subject from the non-diseased population. Suppose that

$$\mathbf{X}|D = d \sim MVN(\mu_d, \Sigma_d), \;\; d = 0, 1,$$

where Σ_0 and Σ_1 are positive definite. For any given real vector $\mathbf{a} \in R^p$,

$$\mathbf{a}^T\mathbf{X}|D = d \sim N(\mathbf{a}^T\mu_d, Q_d),$$

where $Q_d = \mathbf{a}^T\Sigma_d\mathbf{a}$, for $d = 0, 1$. Let $\Phi(\cdot)$ denote the cumulative distribution function of $N(0, 1)$ and $\Phi^{-1}(\cdot)$ be its inverse function. Let $c(u) = \Phi^{-1}(1-u)$ and $\Delta_\mu = \mu_1 - \mu_0$, then at specificity $(1-u)$, the sensitivity of $\mathbf{a}^T\mathbf{X}$ is equal to

$$F(\mathbf{a}, u) = \Phi\left(\frac{\mathbf{a}^T\Delta_\mu - c(u)\sqrt{Q_0}}{\sqrt{Q_1}}\right).$$

Therefore, for a false positive rate region $(0, t)$ for some predetermined $t \in (0, 1)$, the pAUC of $\mathbf{a}^T\mathbf{X}$ is equal to

$$pAUC(\mathbf{a}) = \int_0^t F(\mathbf{a}, u)du. \tag{1}$$

Similar to AUC, the pAUC has the scale invariant property. For identification purposes, in this study the search for the optimal linear combination vector is

restricted to the hyper-sphere with an unit radius. Let $\mathbf{a}^*$ be such a pAUC maximizer; that is,

$$\mathbf{a}^* = \arg\max_{\mathbf{a}\in E_p} pAUC(\mathbf{a}),$$

where $E_p = \{\mathbf{a} | \, \|\mathbf{a}\| = 1, \mathbf{a} \in R^p\}$.

When the population parameters are unknown, the maximum likelihood estimates (MLEs) are employed in a sample version of the optimization problem. Assume two independent random samples of n_0, n_1 are drawn from the non-diseased and diseased populations, respectively. The estimated mean vectors and covariance matrices are, respectively, denoted as follows: $\hat{\mu}_0$, $\hat{\mu}_1$, and $\hat{\Sigma}_0$, $\hat{\Sigma}_1$. Moreover, let $\hat{\Delta}_\mu = \hat{\mu}_1 - \hat{\mu}_0$ and $\hat{Q}_d = \mathbf{a}^T \hat{\Sigma}_d \mathbf{a}$, for $d = 0, 1$. Replacing the unknown parameters in (1) by their corresponding MLEs defined above, we have a sample version of pAUC below:

$$\widehat{pAUC}_n(\mathbf{a}) = \int_0^t \hat{F}_n(\mathbf{a}, u) du, \tag{2}$$

where

$$\hat{F}_n(\mathbf{a}, u) = \Phi\left(\frac{\mathbf{a}^T \hat{\Delta}_\mu - c(u)\sqrt{\hat{Q}_0}}{\sqrt{\hat{Q}_1}} \right).$$

Thus, $\mathbf{a}^*$ of the optimal linear combination is estimated by the maximizer of (2):

$$\hat{\mathbf{a}}_n = \arg\max_{\mathbf{a}\in E_p} \widehat{pAUC}_n(\mathbf{a}).$$

The theorem below shows that $\hat{\mathbf{a}}_n$ is a strong consistent estimator of $\mathbf{a}^*$. A sketch of the proof is provided in Appendix.

Theorem 1. *Assume that $pAUC(\mathbf{a})$ in (1) is a continuous function of $\mathbf{a}$ and has a unique maximizer $\mathbf{a}^*$ in E_p. Then $\hat{\mathbf{a}}_n \to \mathbf{a}^*$ with probability 1 as $n_0, n_1 \to \infty$.*

In real applications, some of the coefficients in the best linear combination were found to be nearly zero. Numerically, their corresponding biomarkers might have limited contribution to the combination and thus to the disease prediction. In the following section, we will discuss how to assess the significance of biomarkers obtained in our maximizing procedure in terms of their discriminatory power. The proposed testing procedures will be embedded into our biomarker selection approaches in order to find a compact biomarker set, which consists of only significant biomarkers in disease diagnosis.

3 Hypothesis Testing and Biomarker Selection

3.1 Testing the Discriminatory Power

Considering only the class of linear combinations, we evaluate the global discriminatory power of a set of $p \geq 1$ biomarkers, $\mathbf{X}$, by testing the following hypotheses:

$$H_{0,g} : \textit{The biomarker set has no discriminatory power to the disease}$$

versus

$$H_{1,g} : \textit{The biomarker set has a discriminatory power to the disease.}$$

The null hypothesis $H_{0,g}$ is true if and only if the optimal linear combination of the biomarker set has no discriminatory power. Or equivalently, the maximal pAUC that the set can achieve through its linear combinations is not greater than the reference limit $t^2/2$, which is the pAUC value of the non-informative diagnosis with a diagonal ROC curve. That is,

$$H_{0,g} : pAUC(\mathbf{a}^*) \leq \frac{t^2}{2} \;\; versus \;\; H_{1,g} : pAUC(\mathbf{a}^*) > \frac{t^2}{2}.$$

By maximizing the sample pAUC defined in (2), we obtain the maximal sample pAUC and use it as the test statistic. That is,

$$T_g = \max_{\mathbf{a} \in E_p} \widehat{pAUC}_n(\mathbf{a}) = \widehat{pAUC}_n(\hat{\mathbf{a}}_n) = \int_0^t \Phi\left(\frac{\hat{\mathbf{a}}_n^T \hat{\Delta}_\mu - c(u)\sqrt{\hat{Q}_0}}{\sqrt{\hat{Q}_1}} \right) du.$$

The null hypothesis $H_{0,g}$ is rejected if T_g is sufficiently large.

Let $\mathbf{X}^T = (\mathbf{X}_{i^-}^T, X_i)$, we consider to assess the contribution of X_i given the existence of other biomarkers, $\mathbf{X}_{i^-}$. The following hypothesis is tested:

$$H_{0,c} : \textit{Given } \mathbf{X}_{i^-}, \textit{ } X_i \textit{ has no discriminatory power to the disease}.$$

The coefficients of the optimal linear combination of $\mathbf{X}$ are written as $\mathbf{a}^{*T} = (\mathbf{a}_{i^-}^{*T}, a_i^*)$, where a_i^* is the corresponding coefficient of X_i. In this problem, we propose evaluating the biomarker X_i from a_i^*. That is, $H_{0,c}$ is equivalent to

$$H_{0,c} : a_i^* = 0.$$

The test statistic is the estimator of a_i^*, denoted by $T_{c,i} = \hat{a}_{n,i}$. The null hypothesis $H_{0,c}$ is then rejected if $T_{c,i}$ is either too small or too large.

Due to the complex formulation of the test statistics, the null distribution and the critical values are estimated by a parametric bootstrapping method. Under the

null hypothesis, the sampling distribution of the test statistic is estimated. Consider drawing two independent random samples of size n_1 and n_0 from the estimated null distribution. Then using the bootstrap samples to find the test statistic. Repeat the sampling B times. The critical value(s) is(are) then equal to the correspondent percentile(s) among these values.

3.2 Biomarker Selection

We now turn to the biomarker selection problem. Assume that all biomarkers are adequately standardized a priori and denoted the full standardized biomarker set by $\mathbf{X}$. Let $\hat{\mathbf{a}}_n^T = (\hat{a}_{n,1}, \ldots, \hat{a}_{n,p})$ be the estimate of the optimal linear combination as before. The magnitude of $|\hat{a}_{n,i}|$ is used as an ordering criterion in the following stepwise biomarker selection approaches. Rearrange the biomarkers according to their corresponding $|\hat{a}_{n,i}|$ values in an ascending order. Denoted the rearranged vector by $\mathbf{X}^T = (X_{(1)}, \ldots, X_{(p)})$.

We consider two stepwise selection methods: the Forward and the Backward approaches. Define A as the set of biomarkers under consideration in each step for convenience. The Forward procedure starts from a null A and tests the contribution of the potentially most discriminatory biomarker $X_{(p)}$. The biomarker is added to A if it is significant. Then it consecutively assesses $X_{(p-1)}$, $X_{(p-2)}$ and so on. On the other hand, the Backward procedure starts from testing the overall discriminatory power of $A = \{\mathbf{X}\}$. If an insignificance is obtained, we stop the selection and conclude that the full biomarker set is independent of the disease. With a significant global effect, one further determines whether the potentially least discriminatory biomarker $X_{(1)}$ is significant. Remove the biomarker from A if an insignificant result is present. Given the result, this procedure consecutively assesses the conditional contribution of $X_{(2)}$, of $X_{(3)}$, and so on. After evaluating the contribution of every individual biomarker, we conclude that the biomarkers remaining in A have a significant contribution to the linear combination in terms of pAUC. The details are presented below:

Forward Method

Step 1. Set $A = \emptyset$. Test the marginal effect of $X_{(p)}$ with respect to

$$H_{0,(p)} : X_{(p)} \textit{ has no discriminatory power.}$$

If $H_{0,(p)}$ is rejected, add $X_{(p)}$ to A. Go to the next step.

Step 2. Test the significance of $X_{(p-1)}$ with respect to

$$H_{0,(p-1)} : \textit{Given } A, X_{(p-1)} \textit{ has no discriminatory power.}$$

If $H_{0,(p-1)}$ is rejected, add $X_{(p-1)}$ to A. Go to the next step.

$\vdots$

Step p. Test the significance of $X_{(1)}$ with respect to

$$H_{0,(1)} : \textit{Given } A, X_{(1)} \textit{ has no discriminatory power.}$$

If $H_{0,(1)}$ is rejected, add $X_{(1)}$ to A. Stop.

Backward Method

Step 0. Set $A = \{\mathbf{X}\}$. Test the global effect of A with respect to

$$H_{0,(0)} : A \textit{ has no discriminatory power.}$$

If $H_{0,(0)}$ is rejected, go to the next step; otherwise, stop and conclude $A = \emptyset$.

Step 1. Assess $X_{(1)}$ by removing $X_{(1)}$ from A and test the hypothesis,

$$H_{0,(1)} : \textit{Given } A, X_{(1)} \textit{ has no discriminatory power.}$$

If $H_{0,(1)}$ is rejected, add $X_{(1)}$ to A. Go to the next step.

Step 2. Assess $X_{(2)}$ by removing $X_{(2)}$ from A and test the hypothesis,

$$H_{0,(2)} : \textit{Given } A, X_{(2)} \textit{ has no discriminatory power.}$$

If $H_{0,(2)}$ is rejected, add $X_{(2)}$ to A. Go to the next step.

$\vdots$

Step p. Assess the effect of $X_{(p)}$. If $A = \{X_{(p)}\}$, stop; otherwise, remove $X_{(p)}$ from A and test the following null hypothesis,

$$H_{0,(p)} : \textit{Given } A, X_{(p)} \textit{ has no discriminatory power.}$$

If $H_{0,(p)}$ is rejected, add $X_{(p)}$ to A. Stop.

Note that except in the initial step of the Backward method, there is no early stopping criterion in both approaches in order to minimize the risk of not taking the variation of $|\hat{a}_{n,i}|$ into the ordering criterion at the beginning of the procedure. Note that at every step, the biomarker set involved is likely to differ, thus the optimal linear combination should be recalculated for the hypothesis testing. Further, the two biomarker selection approaches assess the conditional discriminatory power of one target biomarker at every step and hence the related null hypothesis is $H_{0,c}$. For simplicity, one can consider a fixed significance level stepwisely in the procedure. To control the global type I error rate, a suitable multiplicity adjustment can be employed. For example, the Bonferroni adjustment suggests a α/p stepwise significance level in the Forward method, and $\alpha/(p+1)$ level in the Backward method for the global type I error rate to be controlled at α level.

4 Applications to Real Data Sets

We apply our procedures to two real examples in [6, 19]. By using the raw data and the standardized data, the optimal linear combinations of the full biomarker set are both reported in Table 1. We consider the following standardization: every biomarker in the raw data subtracts the non-diseased group mean and then divides by its pooled sample standard deviation from the two groups for a more uniform unit across biomarkers. The two proposed biomarker selection methods with 5% stepwise significance level are applied on the standardized data. The optimal linear combinations of the reduced biomarker set are given in Table 1.

The first example is a study of Duchenne Muscular Dystrophy (DMD).The DMD carriers generally are elevated by certain serum enzymes, not by physical symptoms. The measurements of three biomarkers of DMD of 87 normal and 38 carrier females were collected in this data set. The sample means of the three biomarkers in the normal and carrier groups are, respectively, $\hat{\mu}_0^T = (3.3932, 4.5213, 2.4863)$, $\hat{\mu}_1^T = (4.7615, 4.5228, 3.0105)$; and the sample covariance matrices are

$$\hat{\Sigma}_0 = \begin{pmatrix} .0316 & -.0039 & .0024 \\ -.0039 & .0065 & .0006 \\ .0024 & .0006 & .0113 \end{pmatrix}, \quad \hat{\Sigma}_1 = \begin{pmatrix} .7683 & -.0050 & .3054 \\ -.0050 & .0094 & -.0064 \\ .3054 & -.0064 & .2268 \end{pmatrix}.$$

We observe from Table 1 that the contribution of the second biomarker is greatly downsized by the standardization. In fact, we find that the marginal distributions of the second biomarker of the two groups do not vary much. Consequently, it should have a limited discriminatory power. The reason that it has an inflated coefficient in the optimal linear combination based on the raw data is due to the fact that it has relatively small variances, which means that it is measured by a greater unit than other biomarkers. The standardization makes the units of the biomarkers more uniform. It leads to a more fair comparison across the biomarkers. After data

Table 1 The estimated best linear combination and the corresponding pAUC for the specificity range (0.9,1) in DMD and atherosclerotic coronary heart disease examples.

Case	Method	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$		$\widehat{pAUC}_n$
DMD	Full set (raw)	0.8350	0.5116	0.2026		0.0888
	Full set (Standardized)	0.9895	0.0653	0.1292		0.0888
	Forward (Standardized)	0.9657	0.0000	0.2597		0.0885
	Backward (Standardized)	0.9657	0.0000	0.2597		0.0885
Case	Method	lutein	TBARS	HDL_C	U_A	$\widehat{pAUC}_n$
Heart disease	Full set (raw)	0.9447	0.3258	0.0265	0.0274	0.0165
	Full set (Standardized)	0.7079	0.6754	0.0834	0.1890	0.0165
	Forward (Standardized)	1.0000	0.0000	0.0000	0.0000	0.0099
	Backward (Standardized)	1.0000	0.0000	0.0000	0.0000	0.0099

standardization, Table 1 shows that both Forward and Backward approaches select the first and the third biomarkers. We find that the decrement in the pAUC by removing the second biomarker is slim.

In another real example, we consider four biomarkers lutein, TBARS, HDL cholesterol (HDL_C), and uric acid (U_A) for construction of a classification tool for atherosclerotic coronary heart disease. A cohort of 434 subjects were selected for the analysis yielding 72 cases and 362 controls. One obtains an insignificant conclusion in testing the null hypothesis of normality. For the non-diseased and diseased groups, the estimated means of the four markers are $\hat{\mu}_0^T = (0.1275, 0.8845, 4.0766, 6.7724)$, $\hat{\mu}_1^T = (0.1402, 0.9337, 4.1225, 6.9112)$; and the two sample covariance matrices are

$$\hat{\Sigma}_0 = \begin{pmatrix} .0034 & -.0004 & -.0002 & -.0051 \\ -.0004 & .0285 & .0039 & .0417 \\ -.0002 & .0039 & .0488 & .0268 \\ -.0051 & .0417 & .0268 & .2846 \end{pmatrix}, \quad \hat{\Sigma}_1 = \begin{pmatrix} .0043 & .0033 & .0006 & .0067 \\ .0033 & .0415 & .0019 & .0426 \\ .0006 & .0019 & .0389 & .0010 \\ .0067 & .0426 & .0010 & .1504 \end{pmatrix}.$$

From Table 1, the impact of the first biomarker lutein, which has relatively small variances in the raw data, is downsized by the standardization. Before the biomarker selection, the first two biomarkers, lutein and TBARS, seem important to the disease from the magnitudes of their coefficients. However, the two stepwise selections produce the same conclusion that only the biomarker lutein achieves statistical significance, although there is a moderate reduction in the pAUC by discarding other three biomarkers.

5 Discussion

In this study, we focus on disease diagnosis with the presence of multiple biomarkers. We consider the class of linear combinations for an effective and easy-to-interpret summarization of the multiple biomarkers. The diagnostic power of a linear combination is evaluated upon its pAUC over a clinically relevant threshold region. In specific, we consider the requirement of a high specificity for the purpose of population screening.

Under the binormality assumption, the pAUC of a linear combination is estimated by employment of MLEs of the population parameters. In addition, the strong consistency of the estimated optimal linear combination is proved. We introduce a testing procedure to assess the overall diagnostic power of a set of biomarkers from the greatest pAUC it can achieve in the class of linear combinations. Furthermore, a testing procedure for determining the conditional contribution of a single biomarker given the existence of other biomarkers is developed. The parametric bootstrap method is applied to find the critical value(s) of the tests. These proposed tests are then embedded in two biomarker selection approaches. The applicability of the proposed methods is illustrated by two real data sets.

Differing with other algorithm-based marker selection approaches, the proposed methods select or discard a biomarker based upon the evidence of statistical significance. As a trade-off, to acquire statistical evidence, our methods necessarily involve many computations. As such, it decreases the feasibility of these methods for big data sets. Consequently, our methods are less appropriate in an exploratory study. We suggest the application of adequate data filtering for dimension reduction prior to advanced statistical confirmatory analysis, such as the construction of a diagnostic rule.

Appendix

Proof of Theorem 1. Since $E(\mathbf{X}|D)^2 < \infty$, by SLLN, as $n \to \infty$,

$$\hat{\mu}_0 \overset{a.s.}{\to} \mu_0,\ \hat{\mu}_1 \overset{a.s.}{\to} \mu_1,\ \hat{\Sigma}_0 \overset{a.s.}{\to} \Sigma_0,\ \hat{\Sigma}_1 \overset{a.s.}{\to} \Sigma_1.$$

Consequently, for any fixed $\mathbf{a} \in E_p$,

$$\mathbf{a}^T\hat{\Delta}_\mu \overset{a.s.}{\to} \mathbf{a}^T\Delta_\mu,\ \hat{Q}_0 \overset{a.s.}{\to} Q_0,\ \hat{Q}_1 \overset{a.s.}{\to} Q_1,\ \text{and}\ \hat{F}_n(\mathbf{a},u) \overset{a.s.}{\to} F(\mathbf{a},u),$$

since $\Phi(\cdot)$ is a continuous function ([18], Theorem 1.10 (i)).

Further, since $\Phi(\cdot)$ is bounded, by the dominated convergence theorem,

$$\lim_{n\to\infty} \widehat{pAUC}_n(\mathbf{a}) = \lim_{n\to\infty} \int_0^t F_n(\mathbf{a},u)du = \int_0^t F(\mathbf{a},u)du = pAUC(\mathbf{a}).$$

Hence, for any fixed $\mathbf{a} \in E_p$, $\widehat{pAUC}_n(\mathbf{a}) \to pAUC(\mathbf{a})$ with probability 1. Since $\mathbf{a}^* = \arg\max_{\mathbf{a}\in E_p} pAUC(\mathbf{a})$ is assumed unique, it implies that for every $\epsilon > 0$, $\sup_{a:d(a,\mathbf{a}^*)\geq\epsilon} pAUC(\mathbf{a}) < pAUC(\mathbf{a}^*)$. As a consequence, according to Theorem 5.7 ([21]),

$$\hat{\mathbf{a}}_n \to \mathbf{a}^*,\ \text{with probability 1.}$$

References

[1] Etzioni, R., Kooperberg, C., Pepe, M., Smith, R. and Gann, P. H. (2003). Combining biomarkers to detect disease with application to prostate cancer. *Biostatistics* **4**, 523-538.

[2] Hsu, M.-J. and Hsueh, H.-M. (2012). The linear combinations of biomarkers which maximize the partial area under the ROC curves. *Computational Statistics*, to appear. DOI: 10.1007/s00180-012-0321-5.

[3] KOMORI, O. AND EGUCHI, S. (2010). A boosting method for maximizing the partial area under the ROC curve. *BFC Bioinformatics* **11**, 314-330.
[4] LASKO, T. A., BHAGWAT, J. G., ZOU, K. H. AND OHNO-MACHADO L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics* **38**, 404-415.
[5] LIN, H., ZHOU, L., PENG, H. AND ZHOU, X.-H. (2011). Selection and combination of biomarkers using ROC method for disease classification and prediction. *The Canadian Journal of Statistics* **39**, 324-343.
[6] LIU, A., SCHISTERMAN, E. F. AND ZHU, Y. (2005). On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in Medicine* **24**, 37-47.
[7] MA, S. AND HUANG, J. (2005). Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* **21**, 4356-4362.
[8] MA, S. AND HUANG, J. (2007). Combining multiple markers for classification Using ROC. *Biometrics* **63**, 751-757.
[9] MADU, C. O. AND LU, Y. (2010). Novel diagnostic biomarkers for prostate cancer. *Journal of Cancer* **1**, 150-177.
[10] MARROCCO, C., DUIN, R. P. W. AND TORTORELLA, F. (2008). Maximizing the area under the ROC curve by pairwise feature combination. *Pattern Recognition* **41**, 1961-1974.
[11] National Cancer Institute: PDQ®Prostate Cancer Screening. Bethesda, MD: National Cancer Institute. Date last modified ⟨06/08/2012⟩. Available at: http://www.cancer.gov/cancertopics/pdq/screening/prostate/HealthProfessional/Page3#Section_67
[12] PEPE, M. S. AND THOMPSON, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123-140.
[13] PEPE, M. S., LONGTON, G., ANDERSON, G. L. AND SCHUMMER, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics* **59**, 133-142.
[14] PEPE, M. S., CAI, T. AND LONGTON, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **62**, 221-229.
[15] RICAMATO, M. T. AND TORTORELLA, F. (2011). Partial AUC maximization in a linear combination of dichotomizers. *Pattern Recognition* **44**, 2669-2677.
[16] ROBIN, X., TURCK, N., HAINARD, A., TIBERTI, N., LISACEK, F., SANCHEZ, J.-C. AND MULLER, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BFC Bioinformatics* **12**, 77-84.
[17] SU, J. Q. AND LIU, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350-1355.
[18] SHAO, J. (1999). Mathematical Statistics. Springer-Verlag Inc.
[19] TIAN, L. (2010). Confidence interval estimation of partial area under curve based on combined biomarkers. *Computational Statistics & Data Analysis* **54**, 466-472.
[20] TURCK, N., VUTSKITS, L., SANCHEZ-PENA, P., ROBIN, X., HAINARD, A., GEX-FABRY, M., FOUDA, C., BASSEM, H., MUELLER, M., LISACEK, F., PUYBASSET, L. AND SANCHEZ, J.-C. (2010). A multiparameter panel method for outcome prediction following aneurysmal subarachnoid hemorrhage. *Intensive Care Medicine* **36**, 107-115.
[21] VAART, A. W. VAN DER (1998). *Asymptotic Statistics.* Cambridge University Press.
[22] WANG, Z. AND CHANG, Y.-C. I. (2011). Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics* **12**, 369-385.
[23] ZHOU, X. H., CHEN, B., XIE, Y. M., TIAN, F., LIU, H. AND LIANG, X. (2012). Variable selection using the optimal ROC curve: An application to a traditional Chinese medicine study on osteoporosis disease. *Statistics in Medicine* **31**, 628-635.

Part III
Bayesian Statistics in Clinical Trials

Safety Concerns of the 3+3 Design: A Comparison to the mTPI Design

Yuan Ji and Sue-Jane Wang

Abstract The 3+3 design is the most common choice by clinicians for phase I dose-escalation oncology trials. In recent reviews, more than 90 % of phase I trials are based on the 3+3 design (Rogatko et al., Journal of Clinical Oncology 25:4982–4986, 2007). The simplicity and transparency of 3+3 allows clinicians to conduct dose escalations in practice with virtually no logistic cost, and trial protocols based on 3+3 pass IRB and biostatistics reviews briskly. However, the performance of 3+3 has never been compared to model-based designs under simulation studies with matched sample sizes. In the vast majority of statistical literature, 3+3 has been shown to be inferior in identifying the true MTD although the sample size required by 3+3 is often magnitude smaller than model-based designs. In this paper, through comparative simulation studies with matched sample sizes, we demonstrate that the 3+3 design has higher risks of exposing patients to toxic doses above the MTD than the mTPI design (Ji et al., Clinical Trials 7:653–663, 2010), a newly developed adaptive method. In addition, compared to mTPI, 3+3 does not provide higher probabilities in identifying the correct MTD even when the sample size is matched. Given the fact that the mTPI design is equally transparent, simple and costless to implement with free software, and more flexible in practical situations, we highly encourage more adoptions of the mTPI design in early dose-escalation studies whenever the 3+3 design is also considered. We provide a free software to allow direct comparisons of the 3+3 design to other model-based designs in simulation studies with matched sample sizes.

Y. Ji (✉)
Center for Clinical and Research Informatics, NorthShore, University HealthSystem, 1001 University Place, Evanston, IL 60201, USA
e-mail: yji@northshore.org

S.-J. Wang
Office of Biostatistics/Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD 20993, USA
e-mail: suejane.wang@fda.hhs.gov

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_11,

1 Introduction

Phase I oncology trials aim to find the maximum tolerated dose (MTD), the highest dose with toxicity rate close to a pre-specified target level, p_T. The 3+3 design [3, 4] is the leading method for phase I dose-escalation trials in oncology, as over 90 % of published phase I trials have been based on 3+3 for the past two decades [1, 5, 6]. Such popularity of 3+3 is striking since numerous model-based dose-escalation methods have been developed by biostatisticians during the same time period and almost all the new methods seemed to exhibit better performance than 3+3 [7–10].

The main reason for the popularity of the 3+3 design is due to its simplicity, transparency, and the costless implementation in practice. In contrast, it often requires a considerable amount of logistic support and complexity to implement most model-based designs. Even if the practical burden could be overcome, protocols based on model-based designs are often subject to more thorough reviews by IRB or among biostatisticians, as operating characteristics of these new designs are required. To the contrary, if the protocol is based on the 3+3 design, such requirement disappears since 3+3 has been widely used. As a result, despite the acceleration in the research development of adaptive model-based designs, the lower standard in the review process and cost-free implementation in practice makes 3+3 an increasingly popular design to physicians. Setting aside the logistic issues, we ask exactly how much better the model-based designs are than 3+3. In reviewing the statistical literature on phase I adaptive designs, we found that when comparing to 3+3, most works did not match the sample size across the designs. For example Ji et al. (2010) [2] showed that 3+3 exhibits a smaller average sample size in the computer simulations than model-based designs, and consequently 3+3 also yields a smaller percentage in identifying the true MTD in these simulations. Since the sample size is not matched in the comparison, it is difficult to assess the reason for the reduced percentage under 3+3. More importantly, since phase I trials focus on patient safety, comparisons without matching sample size cannot provide accurate assessment on the safety characteristics of designs. In fact, usually designs resulting in larger sample sizes should be safer since patients enrolled in the later stage of the trial with a larger sample size will be better protected due to more precise statistical inference.

In this paper, we construct a comprehensive simulation study to evaluate the operating characteristics of 3+3 and a newly developed adaptive design known as the modified toxicity probability interval (mTPI) method [2, 6]. In doing so we match the sample size between the two designs. The main intent of choosing the mTPI design for comparison is because mTPI is equally simple, transparent, and costless to implement. In other words, the logistic burden of mTPI and 3+3 is comparable, which allows us to focus on the simulation performance. Albeit being recently introduced to the society, mTPI has already received attention from both research and industry entities [11, 12]. For example, through personal communication we are informed that almost all phase I oncology trials conducted

at Merck Co., Inc. in the past 2 years have been based on the mTPI design or its variations. Recently, phase I trials based on the mTPI design has been published [13, 14]. Considering the short time period since the publication of the mTPI design, this popularity is encouraging.

In a nutshell, the 3+3 design consists of a set of deterministic rules that dictate dose-escalation decisions based on observed patient outcomes. For example, if out of three treated patients 0, 1, or more than 1 toxicities are observed, 3+3 will recommend escalating dose level, continuing at the same dose level, or de-escalating dose level, respectively (see, e.g., [15, 16]).

The mTPI design uses a Bayesian statistics framework and a beta/binomial hierarchical model to compute the posterior probability of three intervals that reflect the relative distance between the toxicity rate of each dose level to the target rate p_T. Let p_d denote the probability of toxicity for dose d, $d = 1, \ldots, D$, where D is the total number of candidate doses. Using the posterior samples for p_d, mTPI computes the unit probability mass, defined as

$$\text{UPM}_{(a,b)}(d) = \frac{Pr\{p_d \in (a,b) \mid data\}}{b-a}, \tag{1}$$

for three intervals corresponding to *under*-, *proper*-, and *over*-dosing, in reference to whether a dose is lower, close to, or higher than the MTD, respectively. Specifically, the under-dosing interval is defined as $(0, p_T - \epsilon_1)$ and implies that the dose level is lower than the MTD, the over-dosing interval $(p_T + \epsilon_2, 1)$ implies that the dose level is higher than the MTD, and the proper-dosing interval $(p_T - \epsilon_1, p_T + \epsilon_2)$ suggests that the dose level is close to the MTD. Here ϵ_1 and ϵ_2 are small fractions, say 0.05. Inference is robust with respect to the choice of ϵ, as shown in [2]. Large UPM values for each interval imply large per-unit posterior probability mass for that interval, therefore implying the corresponding decision: if $\text{UPM}(d)$ is the largest for under-, proper-, or over-dosing interval, the decision should be to escalate (E), stay (S) at dose d, or de-escalate (D), respectively. Therefore, assuming that dose d is currently used to treat patients, the mTPI design assigns the next cohort of patients based on the decision rule $\mathbf{B}_d$, given by

$$\mathbf{B}_d = \arg \max_{m \in \{D,S,E\}} \text{UPM}(m,d), \tag{2}$$

where $\text{UPM}(m,d)$ is the value of UPM for the dosing interval associated with decision m. Decisions D, S, or E warrant the use of dose $(d-1)$, d, or $(d+1)$ for the next cohort of patients, respectively. Ji et al. [2] proved that the decision rule $\mathbf{B}_d$ is consistent and optimal in that it minimizes the posterior expected loss, in which the loss function is determined to achieve equal prior expected loss for the three decisions, D, S, and E. More importantly, all the dose-escalation decisions for a given trial can be pre-calculated under the mTPI design and presented in a two-way table (Fig. 1). Once the trial starts, clinicians can easily monitor the trial and

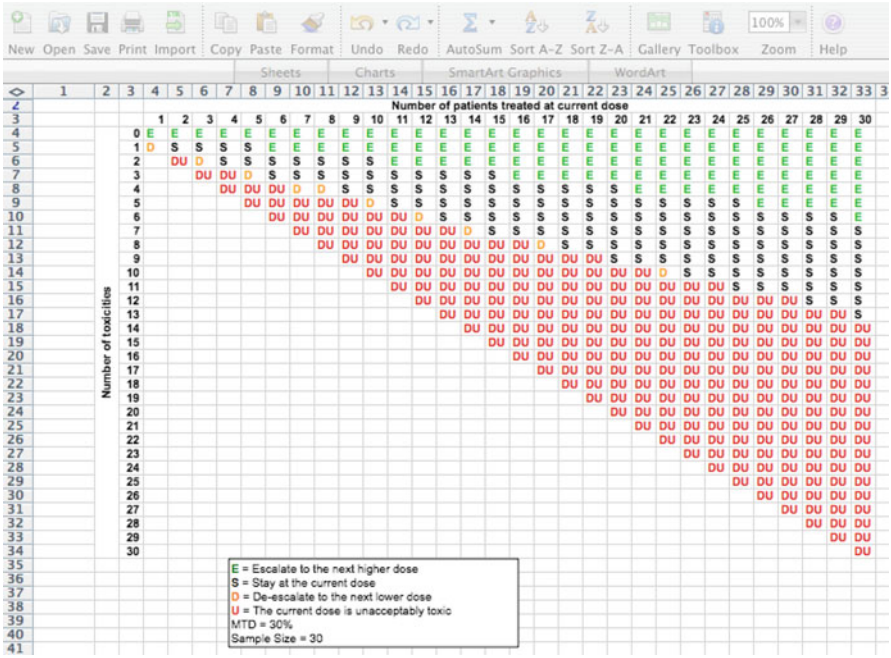


Fig. 1 Dose-finding spreadsheet of the mTPI method. The spreadsheet is generated based on a Beta/Binomial model and pre-calculated before a trial starts. The letters in different colors are computed based on the decision rules under the mTPI method and represent different dose-finding actions. In addition to actions *D*, *S*, and *E*, the table includes action *U*, which is defined as the execution of the *dose exclusion rule* in mTPI.

select the appropriate doses following the pre-calculated table. The simplicity and transparency of mTPI makes it a strong candidate as a model-based counterpart of the 3+3 design in practice. A software in Excel is provided at https://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?Software_Id=72 We will show surprising and important findings and make a recommendation to use mTPI in future phase I trials based on these findings.

2 Comparison of 3+3 and mTPI

2.1 Simulation Setup

We perform computer simulation of phase I trials based on the 3+3 and mTPI designs and compare their operating characteristics summarized over thousands of simulated trials.

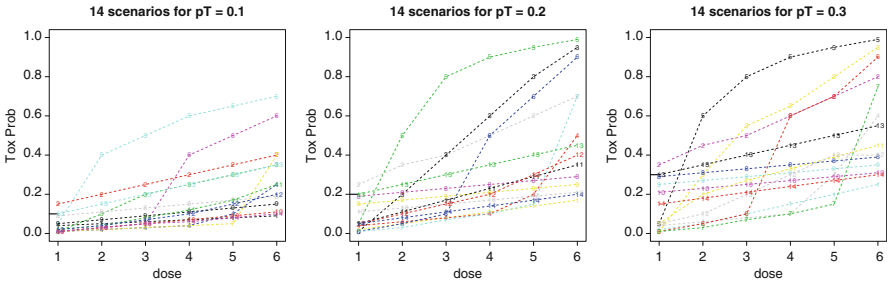


Fig. 2 Dose-response patterns for the 42 clinical scenarios in the simulation. For each of the $p_T = 0.1, 0.2, 0.3$ values, 14 scenarios are constructed.

2.1.1 Clinical Scenarios

We consider 6 doses in the simulated trials. We construct 14 scenarios for each of the three target p_T values, resulting in a total of 42 scenarios. In each scenario, true toxicity probabilities are specified for the 6 doses. These scenarios are set up to capture a wide range of dose–response shapes in practice, as shown in Fig. 2 (see also a discussion in Ji et al., 2012 [17]). Specifically, Scenario 1 represents a case where all doses are safe and low; Scenario 2 represents a case where all doses are high; in Scenarios 3–4 doses cover a wide range of toxicity probabilities and the toxicity probability of one dose equals p_T; Scenarios 5–7 also cover a wide range of toxicity probabilities but the MTD is bracketed by two adjacent doses; In Scenario 8–10, dose toxicity probabilities do not vary much and center around the target p_T; Scenarios 11–12 are similar to Scenarios 8–10, except doses have a wider range of toxicity; lastly, Scenarios 13–14 represent two rare cases in which the MTD is the lowest and highest dose, respectively.

2.1.2 Values of p_T

In practice, the target p_T values are rarely larger than 30 % as it implies unnecessary exposure of patients to doses with high toxicity. Below, we make three choices of p_T: 0.1, 0.2, and 0.3, i.e., the target toxicity rates of the MTD in our simulated trials are 10 %, 20 %, or 30 %. For each p_T and each scenario, we simulate 2,000 trials.

2.1.3 Matching Sample Size

A unique feature in our comparison is that we attempt to match the average sample size of the 3+3 and mTPI designs for each of the clinical scenarios used in the simulation study. To achieve this, for each scenario we first apply the 3+3 design to 2,000 simulated trials and obtain the mean of the 2,000 sample sizes. We then apply

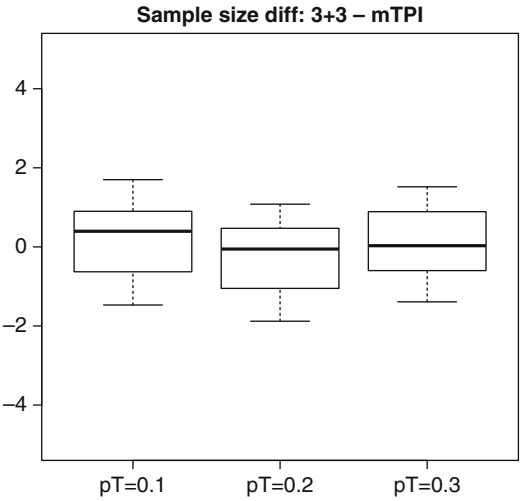


Fig. 3 Difference in the average sample size per trial between 3+3 and mTPI. Each boxplot summarizes the differences for 14 scenarios for a given target toxicity p_T value.

the mTPI design, in which we need to specify the maximum sample size. The mTPI design stops the trial when the total number of patients enrolled is equal or larger than the maximum sample size. We calibrate the maximum sample sizes of mTPI for each p_T value and each scenario, so that the average sample sizes over simulated trials under both designs are similar across all the scenarios. Figure 3 shows the differences of the average sample sizes (over 2,000 simulated trials) between 3+3 and mTPI. The two designs exhibit comparable sample sizes overall. Our calibration of mTPI only involves varying the maximum sample size, while keeping all the other design features unchanged.

2.1.4 Variations of the 3+3

To account for different target p_T values, we use one of the two 3+3 variations (3+3^L and 3++3^H). See Fig. 4. Briefly, the two designs only differ when 6 patients have been treated at a dose, and 1 or 2 of them experience the toxicity. In one variation, 3+3^L, we would stop the trial and declare that the MTD has been exceeded if 2 out of 6 patients experienced toxicity at the dose; in the other variation, called 3+3^H, we would stop the trial and declare that the MTD is that dose. Likewise, 3+3^H would escalate if 1 toxicity is observed from 6 patients, while 3+3^L would stop and declare the dose to be the MTD. Here, ^{L} or ^{H} means that the target toxicity rate p_T of the MTD is low or high. We use 3+3^L for trials with $p_T = 0.1$ or $p_T = 0.2$, and the 3+3^H for trials with $p_T = 0.3$.

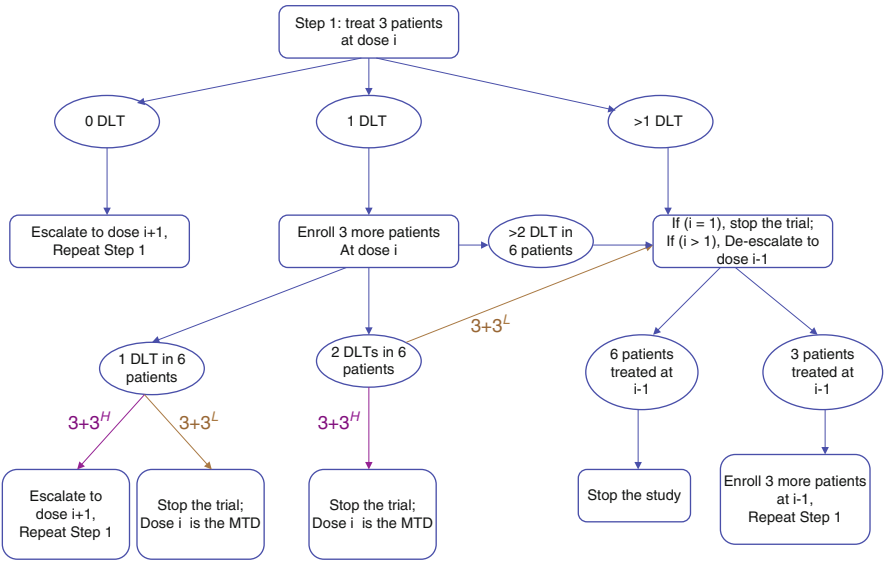


Fig. 4 Schema of the enhanced 3+3 design. The two versions of $3+3^L$ and $3+3^H$ represent the cases where the MTD is defined as the highest dose on which no more than 1 and 2 dose-limiting toxicities (DLT) are observed from 6 patients, respectively.

2.2 *Performance Evaluation*

Summarizing results from 42 scenarios over three different p_T values for three designs can be subjective depending on the criterion used in the comparison. Since the average sample sizes between the two methods are roughly matched, we focus our comparison on two summary statistics simultaneously,

$$n_{>MTD} = \text{ the number of patients treated above the true MTD}$$
$$\%Sel_{MTD} = \text{ the percentage of selecting the true MTD.}$$

$n_{>MTD}$ directly evaluates the safety of each design since under matched sample size; a smaller $n_{>MTD}$ value implies fewer toxicities. To calculate $\%Sel_{MTD}$, we need to decide which doses will be considered as the MTD for each scenario.

2.3 *Main Results*

Figure 5 summarizes the comparison between the 3+3 and mTPI designs, regarding the differences in $n_{>MTD}$ and $\%Sel_{MTD}$. We present the comparison results of $n_{>MTD}$ in the left panel. Comparing to the mTPI design, the 3+3 design has lower

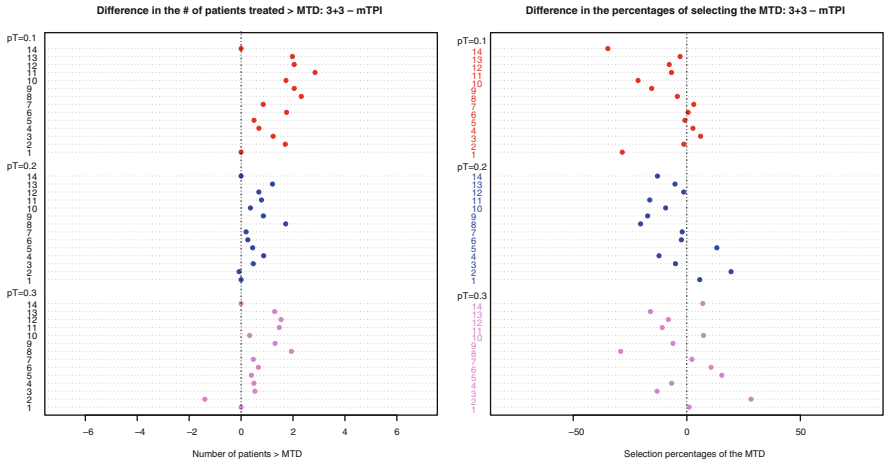


Fig. 5 Comparison between 3+3 and mTPI based on matched sample sizes. The left panel presents the differences in the numbers of patients treated at doses above the MTD ($n_{>MTD}$), i.e., values of ($n_{>MTD3+3}$ - $n_{>MTDmTPI}$) for all 42 scenarios. The right panel presents the differences in the selection percentages of the true MTD ($\%Sel_{MTD}$), i.e., values of ($\%Sel_{MTD3+3}$ - $\%Sel_{MTDmTPI}$) for all 42 scenarios. The three colors in the plots represent the results corresponding to the three different p_T values.

$n_{>MTD}$ values for two scenarios, higher $n_{>MTD}$ for 34 scenarios, and the same $n_{>MTD}$ for six scenarios. In words, 40 out of 42 times, mTPI treats fewer or the same number of patients at doses higher than the MTD than 3+3. In addition, Fig. 6 examines the *overall toxicity percentage*, defined as

$$\frac{\text{the total number of toxicities over all simulated trials}}{\text{the total number of patients treated over all simulated trials}} \times 100\%.$$

Only in one out of 42 scenarios, the 3+3 design exhibits a lower overall toxicity percentage than the mTPI design.

We direct attention to the right panel of Fig. 5 which compares $\%Sel_{MTD}$ between the two designs. In 10 out of 42 scenarios, 3+3 has a higher selection percentage of the true MTD than mTPI. Among these scenarios, the 3+3 design selects the MTD up to about 25 % more often than the mTPI design (Scenario 2 for $p_T = 0.3$). In the remaining 32 scenarios, mTPI selects the MTD more often than 3+3, up to more than 40 % (Scenario 14 for $p_T = 0.1$). A closer examination reveals that 3+3 has higher $\%Sel_{MTD}$ values in scenarios when none of the doses has a toxicity probability close to p_T or when the MTD is at the lower or higher end of the dosing set. We performed additional simulations and confirmed this finding. We found that when the MTD is out of the range of the dosing set, 3+3 usually has a higher selection percentage than mTPI. In other words, 3+3 is a better method when none of the investigational doses is close to the true MTD. This advantage seems to be of limited utility in practice since usually doses are chosen based on scientific and historical data, anticipating some of them are close to the MTD, not the opposite.

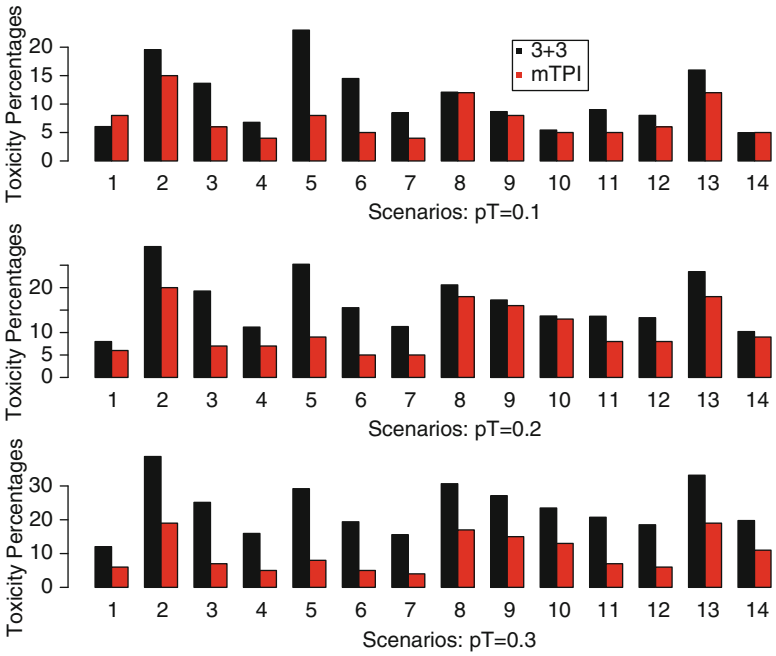


Fig. 6 Overall toxicity percentages for the 3+3 and mTPI designs across all the simulated trials.

Summarizing the two plots in Fig. 5 and considering that (1) the overall sample sizes between the two designs are roughly matched for all the scenarios and (2) the 42 scenarios are constructed to cover a wide range of practical dose–response shapes, *we conclude that the 3+3 design is more likely to treat patients at toxic doses above the MTD and less likely to identify the true MTD than the mTPI design.*

3 Conclusion and Discussion

The mTPI has all the attractive properties 3+3 enjoys for practical considerations and implementations. In addition, compared to the 3+3 design, the mTPI design is safer in treating fewer patients at doses above the MTD, and in general yielding higher probabilities in identifying the true MTD.

In practice, a single value n must be provided as the maximum sample size for the mTPI design in any dose escalation study. In implementing the mTPI design, we recommend a sample size of $n = k \times (d + 1)$ to ensure that the design will reach the highest dose if needed and still has one more cohort to use. Here k is the cohort size and d the number of doses.

It is commonly accepted that phase I trials are of small sizes. This mythology is poorly addressed in the literature. Small phase I trials often provide wrong

recommended doses for phase II, resulting in either low efficacy or high toxicity if the recommended doses are too low or too high, respectively. More discussion and investigation on the proper sample sizes of phase I trials are needed. For example, a streamlined and seamless phase I/II design may result in higher power in the identification of safe and effective doses [18] due to increased sample sizes from the seamless features.

We note that comparison between CRM and 3+3 have been investigated by various authors [19–21] and thus is not included in this paper. A downside of CRM is the lack of easy ways for implementation in practice. We have included the CRM design in our software so that interested users can examine all three designs together, 3+3, CRM, and mTPI.

References

[1] A Rogatko, D Schoeneck, W. Jonas, M. Tighiouart, FR. Khuri, and A. Porter. Translation of Innovative Designs Into Phase I Trials. *Journal of Clinical Oncology*, 25:4982–4986, 2007.

[2] Y Ji, P Liu, Y Li, and BN Bekele. A modified toxicity probability interval method for dose-finding trials. *Clinical Trials*, 7:653–663, 2010.

[3] B.E. Storer. An evaluation of phase I clinical trials designs in the continuous dose-response setting. *Statistics in Medicine*, 48:2399–2408, 2001.

[4] B.E. Storer. Design and analysis of phase I clinical trials. *Biometrics*, 45:925–937, 1989.

[5] C. Le Tourneau, JJ. Lee, and LL. Siu. Dose Escalation Methods in Phase I Cancer Clinical Trials. *Journal of National Cancer Institute*, 101:708–720, 2009.

[6] Y. Ji, Y. Li, and B.N. Bekele. Dose-finding in phase I clinical trials based on toxicity probability intervals. *Clinical Trials*, 4:235–244, 2007.

[7] J. O'Quigley, M. Pepe, and L. Fisher. Continual reassessment method: A practical design for phase i clinical trials in cancer. *Biometrics*, 46:33–48, 1990.

[8] SM Berry, BP Carlin, JJ Lee, and P Müller. *Bayesian Adaptive Methods for CLinical Trials*. CRC, Boca Raton, FL, 2011.

[9] YK Cheung. *Dose Finding by the Continual Reassessment Method*. CRC, Boca Raton, FL, 2011.

[10] B. Neuenschwander, M. Branson, and T. Gsponer. Critical aspects of the Bayesian approach to phase I cancer trials. *Statistics in Medicine*, 27:2420–2439, 2008.

[11] M.S. Blanchard and J.A. Longmate. Toxicity equivalence range design (TEQR): A practical Phase I design. *Comtemporary Clinical Trials*, 32:114–121, 2011.

[12] G.J. Hather and H. Mackey. Some Notable Properties of the Standard Oncology Phase I Design. *Journal of Biopharmaceutical Statistics*, 19:543–555, 2009.

[13] M. Fanale, L. Fayad, B. Pro, F. Samaniego, M. Liboon, C. Nunez, S. Horowitz, P. Anderlini, U. Popat, Y. Ji, LW. Kwak, and A. Younes. Phase I study of bortezomib plus ICE (BICE) for the treatment of relapsed/refractory Hodgkin lymphoma. *British Journal of Haematology*, 154:284–286, 2011.

[14] TA Yap, L Yan, A Patnaik, I Fearen, D Olmos, K Papadopoulos, RD Baird, L Delgado, A Taylor, L Lupinacci, R Riisnaes, LL Pope, SP Heaton, G Thomas, MD Garrett, DM Sullivan, JS de Bono, and AW Tolcher. First-in-Man Clinical Trial of the Oral Pan-AKT Inhibitor MK-2206 in Patients With Advanced Solid Tumors. *Journal of Clinical Oncology*, pages 4688–4695, 2011.

[15] NK Ibrahim, N Desai, and S et al Legha. Phaes I and pharmacokinetic study of ABI-007, a Cremophor-free, protein-stablized, nanoparticle formulation of paclitaxel. *Clinical Cancer Research*, 7(5):1038–1044, 2002.

[16] D Strumberg, H Richly, and RA et al Hilger. Phase I clinical and pharmacokinetic study of the novel Raf kinase and vascular endothelial growth factor receptor inhibitor BAY 43-9006 in patients with advanced refractor solid tumors. *Journal of Clinical Oncology*, 23(5):965–972, 2005.

[17] Y Ji, L Feng, EJ Shpall, P Kebriaei, R Champlin, D Berry, and L Cooper. Bayesian Continual Reassessment Method for Dose-Finding Trials Infusing T Cells with Limited Sample Size. *Journal of Biopharmaceutical Statistics*, 22:1206–1219, 2012.

[18] F Xie, Y Ji, and L Tremmel. A Bayesian adaptive design for multi-dose,randomized,placebo-controlled phase I/II trials. *Contemporary Clinical Trials*, 33:739–748, 2012.

[19] J O'Quigley. Another Look at Two Phase I Clinical Trial Designs. *Statistics in Medicine*, 18:2683–2690, 1999.

[20] E.L. Korn, D. Midthune, T.T. Chen, L.V. Rubinstein, M.C. Christian, and R. Simon. A comparison of two phase I designs. *Statistics in Medicine*, 13:1799–1806, 1994.

[21] S.N. Goodman, M.L. Zahurak, and S. Piantadosi. Some practical improvements in the continual reassessment method for phase I studies. *Statistics in Medicine*, 14:1149–1161, 1995.

Bayesian Interim Inference of Probability of Clinical Trial Success

Ming-Dauh Wang and Grace Ying Li

Abstract Understanding of the efficacy of an investigated compound in early drug development often relies on assessment of a biomarker or multiple biomarkers that are believed to be correlated with the intended clinical outcome. The biomarker of interest may require enough duration of time to show its satisfactory response to drug effect. Meanwhile, many drug candidates in the portfolio of a pharmaceutical company may compete for the limited resources available. Thus decisions based on assessment of the biomarker after a prolonged duration may be inefficient. One solution is that longitudinal measurements of the biomarker be measured during the expected duration, and analysis be conducted in the middle of the trial, so that the interim measurements may help estimate the measurement at the intended time for interim decision making. Considering the small trial size nature of early drug development and convenience in facilitating interim decisions, we applied Bayesian inference to interim analysis of biomarkers.

1 Introduction

Efficiency of drug development has been a focus of discussion in recent years [1]. In the pharmaceutical industry, this has much to do with how a drug company makes critical decisions concerning investment on molecules that will eventually obtain regulatory approval. To make the conventional staged drug development procedure more efficient, drug companies have been seeking for innovative approaches and applying them in the whole drug development process.

One area of attention is on making the early phase drug development more informative so as to render a higher trial success rate of the much more costly

M.-D. Wang • G.Y. Li (✉)
Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46228, USA
e-mail: wang_ming-dauh@lilly.com; li_Ying_Grace@lilly.com

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_12,

late phase confirmatory development. This is in no wise an easy task, in that many characteristics of early phase studies are distinctive from those of the late phase trials. For example, efficaciousness of a drug on treating a disease is normally assessed by measures that require a long treatment and follow-up to show benefits clinically and statistically. Given the exploratory nature of early phase drug development, such efficacy testing is hardly implementable. Thus, biomarkers that are easier to measure and can be therapeutically modified in a relatively shorter period of time would instead be looked upon for making decisions concerning efficacy of the molecule in early drug development. For instance, high density lipoprotein cholesterol (HDL-C) has been the biomarker utilized in early phase programs to determine whether a class of drugs called CETP inhibitors would further reduce cardiovascular events [2, 3]. It takes only a matter of a few months to observe the maximum response of HDL-C to treatment, while it requires trials of a few years long to learn whether a CETP inhibitor can efficaciously reduce cardiovascular risks. Another example is that bone mineral density (BMD) has been used as a biomarker for fracture risks in osteoporosis research [4].

Although biomarkers are relied on to make early drug development decisions, there exist issues that are commonly faced. First, there may not be good biomarkers that have been identified to be well correlated with the intended clinical endpoint. In such a situation, approaches have been proposed to increase the predictive power of biomarkers [5, 6]. Secondly, biomarkers themselves may also need a long duration of time, though still relatively shorter than needed for the clinical endpoint, to exhibit differentiation between a tested drug and its comparator. It is the latter issue that is to be dealt with in our current consideration.

In this manuscript, we consider the situation when early measurements of a biomarker are employed to enhance estimation of the same biomarker at a later time point. This estimation can also involve other biomarkers that bear certain relationships with the primary biomarker of interest. This is done so that at a time during a clinical trial when measurement at the later time point is available only from a small portion of enrolled subjects, informative analysis can be performed to make various critical interim decisions concerning the overall development of the molecule. The framework of Bayesian interim inference of biomarkers in our consideration is laid out in Sect. 2. Section 3 presents an application of the proposed approach. Concluding discussions are given in Sect. 4.

2 Bayesian Interim Assessment of Probability of Trial Success

2.1 The Problem and Bayesian Framework

Consider a clinical trial that is conducted to test the effect of an experimental drug on a biomarker of interest compared to a comparator. The comparator can

be a placebo treatment, an approved agent for the same disease state, or the same compound but in a different delivery form (such as injection versus a non-injection route). To rein in the trial to be of an acceptable size as typical in early phase drug development, the primary measurement of the biomarker is at a prolonged time point that would be regarded as longer than a typical early phase trial. In addition, longitudinal measurements of the biomarker at some earlier time points at and after randomization are taken. The power of the trial is calculated for detection of the anticipated effect size observed with the tested drug at the prolonged time point as compared to the comparator.

Suppose we are interested in the change (or percent change) from randomization (or baseline) in the biomarker, and the following time-response model is assumed:

$$\begin{aligned} Y_1 &= \Delta_1 + \varepsilon = M\left(1 - e^{-kt}\right) + \varepsilon \\ Y_2 &= \Delta_2 + \varepsilon = aM\left(1 - e^{-kt}\right) + \varepsilon, \end{aligned} \tag{1}$$

where Δ_1 and Δ_2 are the mean changes with the comparator and tested drug, M is the mean maximum change at infinity of time, t is the lapse of time from randomization, $k > 0$ is a common shape parameter for both treatments, $a > 0$ is a parameter that measures the relative effect size of the tested drug to the comparator, and $\varepsilon \sim N(0, \sigma^2)$ is the random error.

If multiple doses of the experimental drug are tested, then the model can be expressed as

$$\begin{aligned} Y_i = \Delta_i + \varepsilon &= M_i\left(1 - e^{-kt}\right) + \varepsilon \\ &= a_i M\left(1 - e^{-kt}\right) + \varepsilon, i = 1, \ldots, d, \end{aligned} \tag{2}$$

with $i = 1$; $a_1 = 1$ for the comparator and $i = 2, \ldots, d$; $a_2 > 0, \ldots, a_d > 0$ for the different doses of the tested drug. An alternative approach is to treat dose also as continuous. This can be accomplished by, for instance, the following integrated model:

$$Y = \begin{cases} M\left(1 - e^{-kt}\right) + \varepsilon, & \text{for comparator} \\ \frac{D}{ED50+D}\left[aM\left(1 - e^{-kt}\right)\right] + \varepsilon, & \text{for dose D of tested drug,} \end{cases}$$

where $ED50$ is the dose of the tested drug at which half of the assumed mean maximum effect $aM\left(1 - e^{-kt}\right)$ is achieved when measured at time t.

Another extension is considering within-subject correlations of measurements at the longitudinal time points in the model. Furthermore, other biomarkers that are correlated with the primary biomarker, as well as demographic and genetic variables may also be accounted for in the model to improve evaluation of drug effect on the primary biomarker. For a more extensive treatment of some of these topics, please

refer to [7]. In our experience, modeling issues caused by data scarceness commonly encountered in early phase trials when applying more complex models may suggest falling back to a simpler model.

Longitudinal measurements of change in the biomarker are scheduled to be obtained at times $t = t_k, k = 1,\ldots,K$ from each subject. Denote the time of measurement and change from baseline in measurement of the biomarker at time $t = t_k$ from subject j $(j = 1,\ldots,N_i)$ in treatment i $(i = 1,\ldots,d)$ as t_{ijk} and Y_{ijk}. Then at the interim analysis, all available data $D = (t_{ijk}, y_{ijk})'$s are used to estimate parameters a, M, k, and σ^2 in the model, using model (1) for illustration. In Bayesian inference these parameters are regarded as random variables with a joint prior density denoted by $\pi(a,M,k,\sigma^2)$. Practically, these parameters are assumed independent a priori such that the joint prior density can be written as

$$\pi(a,M,k,\sigma^2) = \pi(a)\pi(M)\pi(k)\pi(\sigma^2).$$

Upon availability of interim data D, information of these parameters is updated by applying the Bayes rule and summarized in the posterior density:

$$\pi(a,M,k,\sigma^2|D) \propto \pi(a,M,K,\sigma^2)\mathcal{L}(a,M,K,\sigma^2|D),$$

where $\mathcal{L}(a,M,K,\sigma^2|D)$ is the likelihood of the parameters given the data.

2.2 Posterior Inference

Interim decisions can be made by drawing information from the posterior distribution concerning comparison of the tested drug with the comparator. For example, non-inferiority of the tested drug in comparison with the comparator at the intended time $t = t_K$ through Model (1) can be assessed by the posterior probability:

$$\begin{aligned} P_1 &= Pr(\Delta_1 - \Delta_2 < \delta|D, t = t_K) \\ &= \int \mathcal{I}\{\Delta_1 - \Delta_2 < \delta|\theta, t = t_K\}\pi(\theta|D)d\theta, \end{aligned} \tag{3}$$

where $\delta > 0$ is the non-inferiority margin, $\mathcal{I}$ is the indicator function, and $\theta = (M,k, a,\sigma^2)$. Non-inferiority is claimed if $P_1 > \eta$, with $0 < \eta < 1$ being the minimally required threshold of probability. At the same time, a low probability below a smaller threshold $0 < \eta' < \eta$ can be evidence for concluding that the tested drug is convincingly inferior to the comparator, and a posterior probability between η' and η is indicative of inconclusiveness.

Another posterior probability not hinging on a single time of measurement, but on the overall trend of the time-response course, is

$$P_2 = Pr(a < \phi|D),$$
$$= \int \mathcal{I}\{a > \phi|\theta\}\pi(\theta|D)d\theta, \quad (4)$$

where $0 < \phi < 1$ is the minimally acceptable discount in effect size by the tested drug to show non-inferiority to the comparator.

When multiple doses of the tested drug are explored as formulated in Model (2), each dose can be compared with the comparator as in (3) or (4). Or if the trial objective is to find whether there is at least one dose that is non-inferior to the comparator, then the following posterior probability would be of interest:

$$P_3 = Pr(\Delta_1 - \Delta_2 < \delta \cup \cdots \cup \Delta_1 - \Delta_d < \delta|D, t = t_K)$$
$$= \int \mathcal{I}\{\Delta_1 - \Delta_2 < \delta \cup \cdots \cup \Delta_1 - \Delta_d < \delta|D, t = t_K\}\pi(\theta|D)d\theta,$$

where θ is extended to be $(M, k, a_2, \ldots, a_d, \sigma^2)$.

Posterior probabilities P_2 and P_3, as P_1, are then compared to selected threshold η and η' for making interim decisions. Computation of P_1, P_2, and P_3 would normally require Monte Carlo simulation, which can be carried out by available Bayesian analysis packages, such as WinBUGS.

2.3 Predictive Inference

The above posterior inference is based on prior information in combination with interim data without prediction. Nevertheless, the yet unobserved measurements at the later time points from the incomplete subjects can be accounted for to make interim predictions for the outcome of the trial if it is continued to the planned end. A yet unobserved value Y at time t can be imputed by the predictive value $\tilde{Y}$ that bears the following density:

$$\tilde{f}(\tilde{y}|D, t) = \int f(y|\theta, t)\pi(\theta|D)d\theta,$$

where $f(y|\theta, t)$ is the density of the measurement at time t given θ. The predicted values are then incorporated in the intended inference.

In the case of testing non-inferiority as discussed in the previous section, the predictive probability of achieving non-inferiority, given the non-inferiority margin δ (or discount in effect size a) and threshold probability η, is calculated by

$$\tilde{P}_1 = \int \mathcal{I}\{Pr(\Delta_1 - \Delta_2 < \delta|\tilde{D}, t = t_K) > \eta\} \prod \left(\tilde{f}(\tilde{y}|D, t)d\tilde{y}\right), \quad (5)$$

$$\tilde{P}_2 = \int \mathcal{I}\{Pr(a < \phi|\tilde{D}) > \eta\} \prod \left(\tilde{f}(\tilde{y}|D, t)d\tilde{y}\right),$$

or

$$\tilde{P}_3 = \int \mathcal{I}\{Pr(\Delta_1 - \Delta_2 < \delta \cup \cdots \cup \Delta_1 - \Delta_d < \delta | \tilde{D}, t = t_K) > \eta\} \prod \left(\tilde{f}(\tilde{y}|D,t)d\tilde{y}\right),$$

where $\tilde{D} = \{D, \tilde{y}'s\}$ are the data that include the predicted values, and the product $\prod$ is of all the yet unobserved values. Again, computation of $\tilde{P}_1$, $\tilde{P}_2$, and $\tilde{P}_3$ typically requires Monte Carlo simulation, such as by WinBUGS.

The predictive probability $\tilde{P}_1$, $\tilde{P}_2$, or $\tilde{P}_3$ is then compared to another pair of thresholds τ and τ', with $0 < \tau' < \tau < 1$. A success is claimed if $\tilde{P}_i > \tau$, and a failure or futility is concluded if $\tilde{P}_i < \tau'$. More in-depth research on predictive inference in interim analysis can be found in [8, 9].

2.4 Interim Decision Making

There are drawbacks and limitations of the conventional early phase decision process, and proposals have been made to make this process more efficient [10, 11]. The purpose of the proposed Bayesian interim analysis is to help facilitate interim decision making. One frequently encountered need is to know if subsequent development activities, such as manufacturing capabilities and toxicology preparation, should be initiated in advance of the end of the ongoing trial. This can be done by examining the interim data, and if the interim results exceed the expectation, then following preparations can be initiated early and accelerated so that unnecessary lags of time between activities can be saved. On the other hand, if the interim results look disappointing, further investment on the molecule can be immediately halted for consideration of patient safety and saving of resources for more promising compounds.

3 An Application

A molecule had been proved for treatment of a certain disease state by a delivery route. Another more convenient delivery route was investigated, which if approved for use was believed to reduce unwanted adverse reactions and thus increase patient compliance and usage. However, this new delivery route needed to show efficacy that is non-inferior to the approved delivery route. A pre-confirmatory Phase 2 trial for testing the new delivery route at 3 doses (Test Dose 1, 2, 3) in comparison with the approved route (Reference) was conducted where the primary efficacy variable was defined as percent change from baseline in a biomarker of interest. This biomarker would take relatively long to fully respond to treatment in the Phase 2 standard. Thus there was the need to assess the biomarker while the trial was ongoing to learn if

1. There is high evidence that the new route is non-inferior to the approved route, and thus full planning and preparations should be at once initiated, or
2. Strength of evidence of the new route being inferior to the old is mediocre, and thus only partial planning be started, or
3. The new route is basically inferior to the approved route, and thus no further planning should be done, but just finish the trial and a comprehensive assessment be conducted to determine subsequent steps.

The new delivery route was believed to behave similarly as the old in terms of time response, but just with a discounted size. The hope was that the discount in effect is minimal so that the impact on the eventual clinical endpoint to be assessed in the confirmatory phase would be acceptable. Considering cost and operation effectiveness, in addition to the measurement of the biomarker at the intended duration of time, only another measurement at an earlier time point was taken to increase power of efficacy detection.

3.1 Prior Elicitation

There were available early phase studies that provide information about the old delivery route. To assess the validity of the assumed model in (1), the nonlinear model was fitted to the data. Both the original means and fitted means are plotted in Fig. 1. As shown in the figure, the model fitted the data reasonably well.

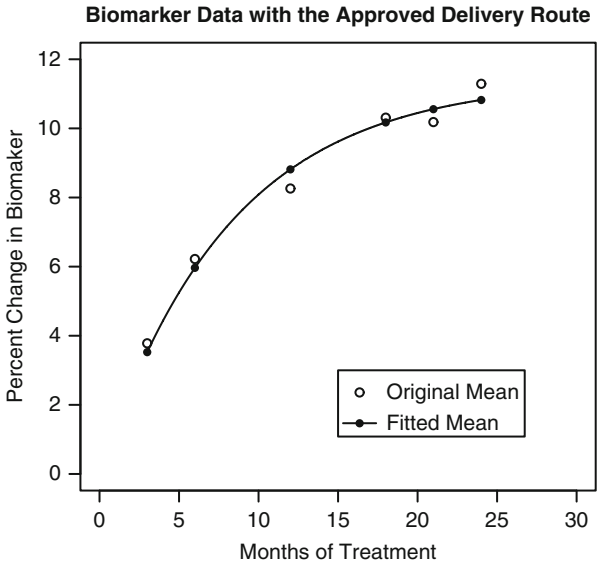


Fig. 1 Data of biomarker of interest.

There was no available data of the biomarker from the new delivery route, yet the prorated model in (1) was deemed reasonable. With this assumption, the data from the approved delivery route as shown in Fig. 1 were summarized to form prior knowledge about the parameters M, k. However, there was no reliable information yet on the effect size of the biomarker with the new delivery route as relative to the old. So a non-informative (or ignorance) prior was assumed for a. A fairly nonspecific weak prior was also assigned for σ^2. Below are the solicited prior distributions:

$$\begin{aligned} M &\sim \text{Gamma}(534, 47), \\ k &\sim \text{Gamma}(65, 523), \\ log(a) &\sim \text{N}(-0.105, 1000), \\ \sigma^2 &\sim \text{Gamma}(0.001, 0.001), \end{aligned}$$

where a gamma distribution $Gamma(\alpha, \beta)$ has mean α/β and variance α/β^2. The gamma distributions for M and k are reflective of the least squares mean estimates of the parameters in the exponential model and their standard errors. The prior mean of -0.105 for $log(a)$ represents an anticipated case of $a = 0.9$ but is diffused by the large variance $1,000$.

3.2 Simulations

Prior to the planned interim analysis, simulations were first conducted to study the performance of the Bayesian inference and likely outcomes. As earlier mentioned, the biomarker was measured at an early time point and at a later time point. It was assumed that at the interim analysis 40 subjects would have measurements at the early time point, and only 10 subjects have measurements at the later time point for each of Reference and Test Dose 1, 2, 3. Non-inferiority is defined with a non-inferiority margin of 3.5 %.

Posterior rather than predictive analysis was implemented for the interim analysis of this trial. Because of the high uncertainty in the likely dose response curve, model evaluation by simulation was done dose-wise for Test Dose 1, 2, 3 in comparison with Reference according to Model (1), whereas the actual interim analysis was conducted using Model (2) that incorporates all treatments. Interim data were first simulated for a range of values of a. The posterior probability that the new delivery route is not inferior to the old at the later time point was then calculated for the different values of a, which are summarized in Table 1. These numbers were used

Table 1 Posterior probability of new delivery route being inferior to the old

a	0.50	0.60	0.65	0.70	0.75	0.80	0.85	0.90
$Pr(\Delta_1 - \Delta_2 < 3.5\% \mid D, t_2)$	0.13	0.28	0.37	0.46	0.56	0.64	0.73	0.80

to guide the selection of interim decision rules. Both the simulation and later on the planned interim analysis were both performed using R and WinBUGS.

3.3 Interim Analysis and Decision

For the purpose of the planned interim analysis, the decision grid tabulated in Table 2 was ascertained across Statistics, Medical, Senior Management, and other disciplines. And this was done with consideration of the simulated results as well as practical probability thresholds for a Phase 2 trial. Because the interim analysis was to decide on the plan for late Phase 2 preparation rather than to potentially stop the trial early for success or futility, the less stringent thresholds 0.35 and 0.65, as compared to 0.2 and 0.8 suggested in [8], were selected. The probability of study success $(PrSS)$ could be the value of the posterior probability in (3) or of the predictive probability in (5). In this application the former was adopted for the interim decision.

At the interim analysis, the time-response model in (2) was employed for the inference of percent change in the biomarker for the reference and 3 doses of the tested delivery route. The posterior results are summarized in Table 3. According to our decision grid, all 3 doses of the tested route failed to pass the futility threshold of 0.2, regardless of safety risk. Thus no late Phase 2 planning was recommended by the interim assessment committee and accepted for implementation.

Upon the futility decision, the trial was not stopped, but continued to the planned end, and thus there was the opportunity to assess the validity of the interim decision. The updated posterior summary given the full trial data was given in Table 4, where similar results are shown as seen at the interim analysis. Therefore, it confirmed the correctness of the futility decision earlier made.

Table 2 Decision grid regarding late phase 2 planning

$PrSS$	Safety Risk		
	Low	Medium	High
$PrSS \geq 0.65$	Full Plan	Partial Plan	No Plan
$0.35 \leq PrSS < 0.65$	Partial Plan	Partial Plan	No Plan
$PrSS > 0.35$	No Plan	No Plan	No Plan

Table 3 Posterior summary of Bayesian interim analysis (percent change from baseline)

Treatment	Mean	SD	HPD (90% Credible Interval)		$Pr(\Delta_1 - \Delta_i < 3.5\% \mid D, t_2)$
Reference	8.48	0.17	8.21	8.78	
Test Dose 1	0.22	0.49	−0.59	1.02	< 0.1
Test Dose 2	1.24	0.50	0.43	2.05	< 0.1
Test Dose 3	1.70	0.47	0.92	2.47	< 0.1

Table 4 Posterior summary of analysis of final data (percent change from baseline)

Treatment	Mean	SD	HPD (90% Credible Interval)		$Pr(\Delta_1 - \Delta_i < 3.5\% \mid D, t_2)$
Reference	8.49	0.18	8.20	8.81	
Test Dose 1	0.35	0.74	−0.83	1.58	< 0.1
Test Dose 2	0.83	0.73	−0.34	2.07	< 0.1
Test Dose 3	1.08	0.70	−0.50	2.22	< 0.1

4 Discussion

Much decreased drug development efficiency in recent years has prompted both regulatory agencies and the pharmaceutical industry to seek ways of reversing the declining trend. These include applying more efficient designs, more powerful analytical methods, and more flexible decision-making processes in clinical trials. Still regarded as non-conventional in drug development and remained much on research shelves not a long while ago, Bayesian inference thus came into scene as showing to be a superior alternative to conventional frequentist approach in many circumstances of pharmaceutical research. In particular, the Bayesian language often proves to be more understandable to non-statisticians and thus helps facilitate communications between statistics and other disciplines.

We have proposed and applied the discussed Bayesian approach for facilitation of interim decisions concerning subsequent activities related to the studied molecule. In our example, this had helped avoid unnecessary further expenditure invested on the seemed futile compound, and thus allowed more promising drug candidates in the portfolio to advance in development with the spared resources. With the intended purpose of the interim analysis, the trial was continued to the planned end after the interim analysis, and thus provided us the opportunity to confirm the correctness of our interim decision.

It is generally recognized that it is not an easy effort to change conventions and cultures that have sticked with drug development for many years. Our example was just a step in educating people in our company on the value of adopting a non-conventional interim decision process during a clinical trial by application of Bayesian inference, with the hope of contributing to the Critical Path Initiative and bearing anticipated fruit.

References

[1] FDA. (2004). Innovation or stagnation? - challenge and opportunity on the critical path to new medical products. http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html.

[2] Cannon, C.P., Shah, S., Dansky, H.M., Davidson, M., Brinton, E.A., Gotto, A.M., Stepanavage, M., Liu, S.X., Gibbons, P., Ashraf, T.B., Zafarino, J., Mitchel, Y., Barter, P. (2010). Safety of anacetrapib in patients with or at high risk for coronary heart disease. *New England Journal of Medicine* **363,** 2406–2415.

[3] Nicholls, S.J., Brewer, H.B., Kastelein, J.J.P., Krueger, K.A., Wang, M.-D., Shao, M., Hu, B., McErlean, E., Nissen, S.E. (2011). Effects of the CETP inhibitor evacetrapib administered as monotherapy or in combination with statins on HDL and LDL cholesterol. *Journal of American Medical Association* **306,** 2099–2109.

[4] Lewiecki, E.M., Miller, P.D., McClung, M.R., Cohen, S.B., Bolognese, M.A., Liu, Y., Wang, A., Siddhanti, S., Fitzpatrick, L.A. (2007). Two-year treatment with denosumab *AMG*162 in a randomized phase 2 study of postmenopausal women with Low BMD. *Journal of Bone and Mineral Research* **22,** 1832–1841.

[5] Chiang, A., Li, G., Ding, Y., Wang, M.-D. (2008). A multivariate ranking procedure to assess treatment effects. *MidWest SAS Users Group - 2008 Conference Proceedings*. http://www.mwsug.org/proceedings/2008/pharma/MWSUG-2008-P04.pdf

[6] Lu, Y., Chiang, A.Y. (2011). Combining biomarker identification and dose-response modeling in early clinical development. *Statistics in Biopharmaceutical Research* **3,** 526–535.

[7] Fu, H., Manner, D. (2010). Bayesian adaptive dose-finding studies with delayed responses. *Journal of Biopharmaceutical Statistics* **20,** 1055–1070.

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

[9] Wang, M.-D. (2007). Sample size reestimation by Bayesian prediction. *Biometrical Journal* **49,** 365–377.

[10] Chuang-Stein, C., Kirby, S., French, J., Kowalski, K., Marshall, S., Smith, M.K., Bycott, P., Beltangady, M. (2010). A quantitative approach for making go/no-go decisions in drug development. *Drug Information Journal* **45,** 187–202.

[11] Brown, M.J., Chuang-Stein, C., Kirby, S. (2012). Designing studies to find early signals of efficacy. *Journal of Biopharmaceutical Statistics* **22,** 1097–1108.

Bayesian Survival Analysis Using Log-Linear Median Regression Models

Jianchang Lin, Debajyoti Sinha, Stuart Lipsitz, and Adriano Polpo

Abstract For the analysis of survival data from clinical trials, the popular semiparametric models such as Cox's (1972) proportional hazards model and linear transformation models (Cheng et al. 1995) usually focus on modeling effects of covariates on the hazard ratio or the survival response. Often, there is substantial information available in the data to make inferences about the median/quantiles. Models based on the median/quantiles (Ying et al. 1995) survival have been shown to be useful in for describing covariate effects. In this paper, we present two novel survival models with log-linear median regression functions. These two wide classes of semiparametric models have many desirable properties including model identifiability, closed form expressions for all quantile functions, and nonmonotone hazards. Our models also have many important practical advantages, including the ease of determination of priors, a simple interpretation of the regression parameters via the ratio of median survival times, and the ability to address heteroscedasticity of survival response. We illustrate the advantages of proposed methods through extensive simulation studies investigating small sample performance and robustness

J. Lin (✉)
Takeda Pharmaceuticals, Cambridge, MA 02139, USA
e-mail: Jianchang.Lin@takeda.com

D. Sinha
Department of Statistics, Florida State University, Tallahassee, Florida 32306, USA
e-mail: sinhad@stat.fsu.edu

S. Lipsitz
Division of General Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA
e-mail: slipsitz@partners.org

A. Polpo
Department of Statistics, Federal University of São Carlos, São Paulo, 13565-905, Brazil
e-mail: polpo@ufscar.br

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_13,

properties compared to competing methods for median regression, which provide further guidance regarding appropriate modeling in clinical trial.

1 Introduction

Semiparametric models such as Cox (1972) proportional hazards model and linear transformation models (Cheng et al. 1995; Fine et al. 1998) are very popular for modeling effects of covariates on survival response. For example, the main aim of a semiparametric model for a two-arm randomized trial for small cell lung-cancer (SCLC) patients (Ying et al. 1995) is to express the effects of treatment arm and age at entry on time to death (survival time). Often, there is substantial data information available about the median. However, previous semiparametric models for survival data do not focus on the effects of covariates on the median and other quantiles. Particularly for Bayesian survival analysis, medians and other quantiles are natural choices for elicitation of experts opinions. Clinical experts on the disease under study are likely to have useful prior information/opinions about survival quantiles (say, the median) and the changes in the median for varying covariate values. However, semiparametric Bayesian models for survival data, possibly with the exception of Kottas and Gelfand (2001), and Hanson and Johnson (2002), are either based on covariate effects on the hazard ratio (see Ibrahim et al. 2001) or on the mean survival time (e.g., Walker and Mallick 1999. In two-arm cancer clinical trials, the determination of a clinically significant difference and subsequent evaluation of power of the trial, even for frequentist trial designs, are often based on the prior evaluation of the median for the control arm as well as the clinically significant effect of treatment on median survival time (Piantadosi 2005). In this paper, our goal is to propose novel semiparametric models for median survival time with interpretable regression effects. We show that these wide classes of semiparametric models have many desirable properties including model identifiability and non-monotone hazards. Unlike previous methods for Bayesian survival analysis (e.g., Hanson and Johnson 2002), our models accommodate the situation when the location/median as well the scale and shape of the survival distribution are affected by the covariate. Unlike some of the previous frequentist methods for median regression, we do not require the restrictive assumption that all quantile functions below the median to be linear.

In Sect. 2, we introduce two new semiparametric survival models with log-linear median regression functions. In Sect. 2, we also show the desirable properties of these two large classes of survival models, including closed form expressions for other quantile functions (besides the median). We also present the comparisons as well as relationships of our models with existing Bayesian and frequentist median regression models. In Sects. 3 and 4, we present the likelihood, suitable nonparametric prior processes and Markov Chain Monte Carlo (MCMC) tools to estimate the model parameters using a semiparametric Bayesian approach. In Sect. 5, our simulation studies reveal that estimators based on our models have better

small sample performances and more robustness properties compared to competing methods for median regressions including the estimators of Portnoy (2003). Some final remarks are in Sect. 6.

2 Semiparametric Models

Let T_i be the survival time of subject $i = 1, \ldots, n$ and let $Z_i = (1, Z_{i1}, \ldots, Z_{ip})'$ be the corresponding vector of p time-constant covariates along with the intercept term. The monotone power transformation $g_\lambda(y)$ (Bickel and Doksum 1981),

$$g_\lambda(y) = \frac{\text{Sgn}(y)\ |y|^\lambda}{\lambda} \text{ for } \lambda > 0\,, \tag{1}$$

where $\text{Sgn}(y) = -1$ for $y < 0$ and $\text{Sgn}(y) = +1$ otherwise, is an extension of the Box-Cox power family (Box and Cox 1964), a popular transformation to obtain symmetric and unimodal density for the transformed random variable. We assume that for unknown λ, the transformed survival time $g_\lambda\{\log(T_i)\}$ is symmetric and unimodal with median $g_\lambda(\beta' Z_i) = g_\lambda(M_i)$, that is,

$$g_\lambda\{\log(T_i)\} = g_\lambda(M_i) + \varepsilon_i, \tag{2}$$

where ε_i are iid from a unimodal and symmetric density $f_\varepsilon(\cdot)$ centered at 0, $M_i = \beta' Z_i$, and β is the vector of regression parameters. Carroll and Ruppert (1984) and Fitzmaurice (2007) among others proposed parametric version of the transform-both-sides (TBS) regression model for an uncensored continuous response with the original Box-Cox transformation (Box and Cox 1964) and $N(0, \sigma^2)$ density for error $f_\varepsilon(\cdot)$.

The transformation $g_\lambda(y)$ in (1) is monotone with derivative $g'_\lambda(y) = |y|^{\lambda-1}$. The median of $\log(T_i)$ is $M_i = \beta' Z_i$ because $P[\log(T_i) > M_i] = P[g_\lambda\{\log(T_i)\} > g_\lambda(M_i)] = F_\varepsilon(0) = 1/2$, where F_ε is the cdf of ε. As a consequence, the survival time T_i has a log-linear median regression function $Q_{0.5}(Z_i) = \exp(M_i) = \exp(\beta' Z_i)$ and survival function $S(t|z) = 1 - F_\varepsilon(g_\lambda(\log t) - g_\lambda(M))$. The following theorem shows that the parameter λ and the density f_ε of (2) are also identifiable, in the sense that for any survival time following (2), there is a unique (λ, f_ε) for which $g_\lambda\{\log(T_i)\}$ has a symmetric unimodal distribution.

In relation to the parametric TBS model with Gaussian ε, Kettl (1991) introduced another parametric model $g_\lambda\{\log(T_i)\} = g_\lambda(M_i) + |M_i|^\gamma \eta_i$ with $\eta_i \sim N(0, \sigma^2)$. As an alternative to the semiparametric TBS model of (2), we can consider a semiparametric extension of Kettl's (1991) model with a symmetric unimodal density for η_i. However, for the sake of parsimony, we recommend not using two separate parameters λ and γ to achieve symmetry and to model heteroscedasticity, respectively. Instead, we suggest the Power-Of-Mean-Scale (POMS) semiparametric model

$$\log(T_i) = M_i + |M_i|^{\gamma}\varepsilon_i \tag{3}$$

with $\varepsilon_i \sim f_\varepsilon(.)$ symmetric and unimodal at 0, as an alternative to the TBS model. The key assumption of the semiparametric POMS model of (3) is that $\log(T)$ is symmetric and unimodal with median $M_i = \beta Z_i$. Unlike the Bayes model of Kottas and Gelfand (2001) and Hanson and Johnson (2002), the POMS model of (3) takes care of the heteroscedasticity of $\log(T_i)$. The survival function of (3) is $S(t|Z) = S_\varepsilon\{\log(t^{|M|^{\gamma}}e^{-sgn(M)|M|^{1-\gamma}})\}$, and it reduces to the accelerated failure time model with $S(t|Z) = S_\varepsilon\{\log(te^{-M})\}$ when $\gamma = 0$.

Although the models in (2) and (3) apparently focus on modeling the median, we can easily obtain other quantiles of $\log(T)$. For TBS of (2), the α-quantile $Q_\alpha(Z)$ of T is

$$Q_\alpha(Z) = \exp\{M^*_\alpha(Z)\} = \exp\left[g_\lambda^{-1}\{g_\lambda(\beta' Z) + \varepsilon^*_\alpha\}\right], \tag{4}$$

because $P[g_\lambda\{\log(T)\} < g_\lambda(M) + \varepsilon^*_\alpha \ |Z] = \alpha$ for $\alpha \in (0,1)$, where ε^*_α is the α-quantile of $f_\varepsilon(\cdot)$ with $P(\varepsilon < \varepsilon^*_\alpha) = \alpha$. For $\alpha = 0.5$, we have $\varepsilon^*_{0.5} = 0$ and get the log-linear median function $\exp(\beta' Z)$ for T in (2). Similarly, the α-quantile of T for the POMS model is

$$Q_\alpha(Z) = \exp(M + |M|^{\gamma}\varepsilon^*_\alpha), \tag{5}$$

where $M = \beta' Z$ for the regression parameter β in (3). These expressions of (4) and (5) show that both of these models are very convenient for simultaneously estimating all important quantiles of T_i using the estimates of parameters and ε^*_α. However, unlike the existing methods including those of Portnoy (2003) and Peng and Huang (2008), $Q_\alpha(Z)$ of the TBS and POMS models in (4) and (5) are not linear in covariate Z unless $\alpha = 0.5$ (median). The Bayesian models of Kottas and Gelfand (2001) and Hanson and Johnson (2002) also have linear quantile functions $M_\alpha(Z) = \beta'_\alpha Z$ of $\log T$ for all $1 > \alpha > 0$, and they are parallel to each other (with only the intercept of β_α different for different $\alpha \in (0,1)$).

3 Model Estimation and Inference

Let T_i and C_i be the survival and censoring times, respectively, for $i = 1, \cdots, n$. We observe (t_{i0}, δ_i), where $t_{i0} = T_i \wedge C_i$ is the observed follow-up time and δ_i is the censoring indicator, with $\delta_i = 1$ for $T_i = t_{i0}$ and 0 otherwise. It is assumed that T_i and the random censoring time C_i are conditionally independent given covariate Z_i. Given the observed data vector $(\mathbf{t_0}, \delta^*)$ with $\mathbf{t_0} = (t_{10}, \ldots, t_{n0})$ and $\delta^* = (\delta_1, \ldots, \delta_n)$, the likelihood function under our TBS model of (2) is as follows:

$$L(\beta, \lambda, f_\varepsilon | \mathbf{t_0}, \delta^*) \propto \prod_{i=1}^{n} \left\{ |y_i|^{\lambda - 1} f_\varepsilon(\omega_i) \right\}^{\delta_i} \{1 - F_\varepsilon(\omega_i)\}^{1-\delta_i}, \tag{6}$$

where $\omega_i = g_\lambda(y_i) - g_\lambda(\beta' Z_i)$ with $y_i = \log(t_{i0})$, $F_\varepsilon(\omega) = \int_{-\infty}^{\omega} f_\varepsilon(u)du$ is the cdf of the unimodal symmetric density function f_ε. For the POMS model of (3), the corresponding likelihood is

$$L(\beta, \gamma, f_\varepsilon | \mathbf{t_0}, \delta^*) \propto \prod_{i=1}^{n} \left\{ f_\varepsilon(\omega_i^*) |M_i|^{-\gamma} \right\}^{\delta_i} \left\{ \overline{F}_\varepsilon(\omega_i^*) \right\}^{1-\delta_i}, \tag{7}$$

where $\omega_i^* = (y_i - M_i)/|M_i|^\gamma$, $M_i = \beta' Z_i$ and $\overline{F}_\varepsilon(u) = \int_u^{+\infty} f_\varepsilon(z)dz$.

In general, for the parametric versions of either TBS or POMS model, any unimodal symmetric distribution, such as the Gaussian and logistic, can be used for F_ε. For example, $f_\varepsilon(w)$ and $F_\varepsilon(w)$ will be, respectively, replaced by the density $\phi_\sigma(w)$ and cdf $\Phi_\sigma(w)$ of $N(0, \sigma^2)$ for the Gaussian TBS model likelihood in (6) and the Gaussian POMS model's likelihood in (7). The corresponding posterior is $p(\tau, \sigma | \mathbf{t_0}, \delta^*) \propto L(\tau, \sigma | \mathbf{t_0}, \delta^*)\pi(\tau, \sigma)$, where $\pi(\tau, \sigma)$ is the joint prior density based on the available prior information, with $\tau = (\beta, \lambda)$ for TBS and $\tau = (\beta, \sigma)$ for POMS. MCMC samples from this joint posterior can be used to implement a parametric Bayesian analysis. Under either of these two parametric models, the maximum likelihood estimator (MLE) of the regression parameters β can be obtained via maximizing the log-likelihood $L(\tau, \sigma | \mathbf{t_0}, \delta^*)$ corresponding to the chosen model.

4 Semiparametric Bayesian

Any parametric assumption about f_ε is deemed as a restrictive parametric assumption for some data examples in practice. In the semiparametric models of (2) and (3), the unimodal symmetric distribution F_ε of error ε_i is assumed unknown. The semiparametric likelihood of model in (2) is given as

$$L(\beta, \lambda, F_\varepsilon | \mathbf{t_0}, \delta^*) \propto \prod_{i=1}^{n} \left\{ |y_i|^{\lambda - 1} \, dF_\varepsilon(\omega_i) \right\}^{\delta_i} \left[\bar{F}_\varepsilon(\omega_i) \right]^{1-\delta_i}, \tag{8}$$

where $\bar{F}_\varepsilon(u) = 1 - F_\varepsilon(u)$ and $\omega_i = g_\lambda(y_i) - g_\lambda(\mu_i)$. Similarly, the semiparametric likelihood of POMS model of (3) is

$$L(\beta, \gamma, F_\varepsilon | \mathbf{t_0}, \delta^*) \propto \prod_{i=1}^{n} \left\{ |M_i|^{-\gamma} \, dF_\varepsilon(\omega_i^*) \right\}^{\delta_i} \left[\bar{F}_\varepsilon(\omega_i^*) \right]^{1-\delta_i}, \tag{9}$$

where $\omega_i^* = (y_i - M_i)/|M_i|^\gamma$ for $M_i = \beta' Z_i$. For semiparametric maximum likelihood estimation (SPMLE) under these models, the likelihoods are maximized with respect to the restriction that F_ε is the cdf of a unimodal distribution symmetric around 0. The regularity conditions and asymptotic issues for the SPMLE under either (8) or (9) are nontrivial and beyond the scope of this paper. For semiparametric Bayesian analysis, we need the posterior

$$p(\tau, F_\varepsilon|\mathbf{t_0}, \delta^*) \propto L(\tau, F_\varepsilon|\mathbf{t_0}, \delta^*)\pi_{12}(\tau)\pi_3(F_\varepsilon)\ , \tag{10}$$

where π_{12} and π_3 are independent priors of τ and F_ε. We would like to emphasize that the expression of the likelihood $L(\tau, F_\varepsilon|\mathbf{t_0}, \delta^*)$ in (10) and even the priors π_{12} and π_3 actually depend on the underlying semiparametric model (either TBS of (2) or POMS of (3)). This uses the simplifying, however reasonable, assumption that the prior opinions about parametric vector τ and nonparametric function F_ε can be specified independently. We will discuss the practical justification of this assumption later.

We now introduce a class of nonparametric priors π_3 for F_ε, applicable either for model in (2) or in (3), defined over the space of symmetric unimodal distribution functions. We use the result that any symmetric unimodal distribution F_ε can be expressed as a scale-mixture of uniform random variables

$$F_\varepsilon(u) = \int_0^\infty \zeta(u|\theta)\, dG(\theta) \tag{11}$$

for some mixing distribution $G(\theta)$ (Feller 1971 p.158), where $\zeta(u|\theta)$ for $\theta > 0$ is the uniform distribution with mean zero and support $(-\theta, +\theta)$. We use the Dirichlet process (Ferguson 1973) $G \sim DP(G_0, \nu)$ prior for the unknown scale-mixing distribution $G(\theta)$ of (11) to define a nonparametric prior for the random (unknown) unimodal symmetric distribution F_ε. The Dirichlet process (DP) is characterized by the known "prior guess" G_0 (the prior expectation of G), and a positive scalar parameter ν, precision around prior mean/guess G_0. The prior mean G_0 of the random mixing density G can be chosen appropriately to assure a desired prior mean/guess F_* for unknown F_ε. Using a result by Khintchine (1938), when the density $f_*(\cdot)$ and its derivative $f'_*(\cdot)$ exist, the density $G'_0(\theta)$ of $G(\theta)$ is given as

$$G'_0(\theta) = -2\theta\ f'_*(\theta)\ \text{ for } \theta > 0\ . \tag{12}$$

For example, to obtain an approximate double exponential ($Dexpo(\gamma)$) prior mean density $f_*(\varepsilon) = \frac{1}{2}\gamma\exp(-\gamma|\varepsilon|)$ for the regression error density f_ε, using (12), we need to choose $G_0(\theta|\gamma)$ as $Gamma(2, \gamma)$ with density $G'_0(\theta|\gamma) = \gamma^2\theta\exp(-\gamma\theta)$.

5 Simulation Study

For our simulation models, we set the median of $Y = \log(T)$ given Z to be $M(Z) = \beta_0 + \beta_1 Z = 6.5 + Z$ with $\beta_0 = 6.5$ and $\beta_1 = 1.0$, where Z can take four possible values $0, 0.5, 1.0$, and 1.5, in equal proportions for each simulated dataset. For each simulation distribution of T considered in the study, we simulate at least 5,000 datasets with sample sizes $n = 80, 160$, and 320. The number of simulated datasets for different sample sizes may vary to assure that the Monte Carlo variability of the approximate bias and MSE of the regression estimates being smaller than 0.01.

Table 1 Results of simulation study under Exponential and Pareto models: Monte Carlo approximation of the sampling mean and Mean Square Error (MSE) of different estimators of known $\beta_1 = 1$

Simulation		TBS MLE		Portnoy		SP TBS		SP POMS	
Model	Sample	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
Exponential	80	0.91	2.66	0.93	4.27	0.92	0.90	1.06	1.50
	160	0.97	1.35	1.11	2.28	1.08	0.65	1.09	1.12
	320	0.94	0.69	0.96	1.20	0.93	0.48	1.07	0.68
Pareto	80	1.00	12.69	1.12	26.98	1.03	0.95	0.73	1.76
	160	0.99	6.00	1.08	9.96	1.01	0.85	0.86	4.57
	320	0.95	2.75	0.97	4.41	1.02	0.68	0.88	2.23
TBS	80	0.99	1.94	1.01	1.52	1.04	0.72	0.92	0.83
	160	0.96	0.97	0.98	1.69	0.97	0.48	0.92	0.70
	320	0.97	0.51	0.98	1.35	1.03	0.30	0.97	0.72

First we evaluate the robustness of the MLE and of the semiparametric Bayes estimates based on the TBS model of (2) and the POMS model of (3). We compare performances (bias and MSE) of these three estimators to the competing frequentist estimator of Portnoy (2003). For this aim, we simulate data from a parametric conditional distribution of Y given Z, namely the Exponential and Pareto, where the assumptions of TBS in (2) and POMS in (3) are not valid. Both exponential and pareto simulation densities are heteroscedastic and skewed for all λ. The independent censoring was generated from an Exponential density with rate parameter k chosen to obtain desired proportions of censoring. For example, the choice of $k = \log(2)/30$ results in approximately 20% censoring for exponential simulation model.

Table 1 presents the summary of the approximate sampling mean and mean-square-error (MSE) of various competing estimators of β_1 under exponential and Pareto simulation models. Results in Table 1 under an exponential simulation model show that the MLE based on (2) and the Bayes estimators based on either (2) or (3) have comparable biases relative to competing estimators. Further, the MSE of Portnoy's estimators are much larger than the corresponding MSE of the MLE and Bayes estimators. The Bayes estimators under (2) and (3) have much smaller MSE compared to the MLE.

In simulations for the Pareto distribution with scale parameter equal to 1, $g_\lambda(Y)$ is an extremely skewed and heavy-tailed density for all values of λ. As we expect, the bias of the MLE $\widehat{\beta}_1$ is slightly higher than that of Portnoy's estimator. The bias of semiparametric Bayes estimators are similar to the bias of Portnoy's estimator. However, they have much smaller MSE than other competing estimators. For the Pareto simulation model, the MSE of Bayes estimators under the TBS model are substantially smaller compared to that of Bayes estimators based on POMS model.

For the last part of Table 1, our aim is to investigate the improved performance of the Bayes estimator under a correct TBS modeling assumption compared to rest of the competing estimators. For this aim, we simulated data from TBS model of (2) with $\lambda = 0.5$ and double-exponential density for ε. We see that Bayes

estimators under the TBS models have substantial improvement in MSE compared to competing estimators including the Bayes estimators under the POMS model. The bias of MLE under the Gaussian TBS model increases with sample size. The MSE for Bayesian POMS model is worse (30–45 % increase) than that of Bayes TBS model for all sample-sizes.

In summary, when the distribution of $\log T$ after an optimal transformation has a moderate degree of asymmetry, the MLE and Bayes estimators based on either (2) or (3) have finite sample biases very similar to that of Portnoy (2003)'s estimator. More importantly, the precision of Bayes estimators based on TBS and POMS are better even when the underlying assumptions of either (2) or (3) are not entirely valid. However, the MLE's performance depends on the degree of symmetry of the distribution of $g_\lambda(Y)$ under optimal λ. The semiparametric Bayes estimators have excellent biases and smallest MSE among all of its competitors. When the modeling assumption of (2) is correct, the Bayes estimator based on (2) shows much better MSE compared to any competing estimators. This implies that the semiparametric Bayes estimator based on (2) is a safer and more robust estimator to use in practice compared to its competitors.

6 Discussion

In this paper, we present two new classes of semiparametric models amenable to Bayes estimation of the log-linear median regression function for censored survival data. Similar to previous semiparametric models (e.g., Cox's model), both models have one nonparametric function f_ε and a finite dimensional parameter vector. Our Bayes methods have the advantage of the ease of determination of priors and a simple interpretation of the regression parameters via the ratio of median survival times. Our model classes are large since the nonparametric error density f_ε in either model has no assumed functional form except the assumption of symmetry and unimodality. We argue that the assumption of unimodality is needed because it justifies the importance of median as the location parameter of interest. As mentioned in previous literature including Box and Cox (1964), the transformation in (1) is often an effective tool to obtain symmetry and accommodate heteroscedasticity. Our method can be applied when the covariate Z affects the location as well as the scale and shape of the regression error of $\log T$.

In spite of the many similarities in interpretations and forms between the TBS and POMS model, there are some differences between them. The expression in (4) for the TBS model also implies that $Q_\alpha(Z_i) \leq Q_\alpha(Z_j) \Leftrightarrow Q_{\alpha'}(Z_i) \leq Q_{\alpha'}(Z_j)$ for all $\alpha, \alpha' \in (0,1)$. This means that under the model in (2), ordering between two patients' median survival times implies uniform ordering between their corresponding survival functions over the entire time-axis. This property is similar to Cox's model where ordering between two hazards (as well as survival functions) remain the same over the entire time-axis. However, the POMS model of (2) does not satisfy this property. Under the POMS model of (3), survival functions from two treatment

groups may cross each other. The key assumption of symmetric $\log T$ in (3) may be restrictive in some applications. Our simulation study shows that MSE of estimator under (3) is higher than that under (2) when $\log T$ has skewed and heteroscedastic density.

Although we focus on modeling the median functional, our methods can be either used via (4) or (5) to compute even the joint confidence band of any other quantile functional because the quantile functions of (4) and (5) are known functions of $(\beta, \lambda, \varepsilon_\alpha^*)$. For brevity, we have omitted the results of our simulation study showing an excellent accuracy of joint confidence bands of all these quantile functions $(Q_{0.25}(z), Q_{0.5}(z), Q_{0.75}(z))$ under Bayes TBS model of (2) (even when the simulation model is Pareto). However, for some diseases, such as cancers with very good prognosis, the main interest may not be on the median, and the goal may be on modeling the quantile $Q_\alpha(Z)$ as a log-linear function with $P\{T < Q_\alpha(Z)\} = \alpha$ for $\alpha > 1/2$. In this case, we can use a modification of the log-linear models in (2) and (3) with assumption $P(\varepsilon < 0) = F_\varepsilon(0) = \alpha$. We can use the scale-mixture model of (12) with the modification that $\zeta(u|\theta)$ being the uniform density with support $\{2\theta(\alpha - 1),\ 2\theta\alpha\}$ where θ has an unknown mixing density G. For the sake of brevity, we again omitted the details of the rest of the methodology and related MCMC steps. This model allows only the $(1-\alpha)$-percentile of T to be a log-linear function of covariate.

Our methods can also predict the outcome of a future patient with known covariate values. Our simulation results show that the efficiency gain of semiparametric Bayes estimators is substantial compared to existing frequentist estimators even when our assumptions of (2) and (3) do not hold (e.g., for Pareto distribution). We do not present any separate simulation study of parametric Bayes estimators because these estimators under diffuse prior information are numerically close to parametric ML estimators. All of these advantages make our proposed methods extremely attractive alternatives to other existing semiparametric method for censored data.

References

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.

Cheng, S.C., Wei, L.J. and Ying, Z. (1995) Analysis of transformation models with censored data. *Biometrika* **82,** 835–845.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–200.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*. Wiley.

Ferguson, T.S. (1973), Bayesian Analysis of some nonparametric problems, *The Annals of Statistics* **1,** 209–230.

Fine, J.P., Ying, Z. and Wei, L.J. (1998), On the linear transformation model with censored data, *Biometrika* **85**, 980–986.

Fitzmaurice, G.M., Lipsitz, S.R. and Parzen, M. (2007) Approximate median regression via the Box-Cox transformation. *The American Statistician*. **61,** 233–238.

Hanson, T. and Johnson, W.O., (2002), Modeling regression error with a mixture of Polya trees, *J. Amer. Statist. Assoc.* **97**, 1020–1033.

Ibrahim, J.G., Chen, M.-H., Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag.

Khintchine, A.Y. (1938). *On Unimodal Distributions*. *Inst. Mat. Mech. Tomsk. Gos. Univ.* **2,** 1–7.

Kettl, S. (1991). Accounting for heteroscedasticity in the transform both sides regression model. *Applied statistics*, **49**, 261–268.

Kottas, A. and Gelfand, A.E. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, **456**, 1458–1468.

Portnoy, S. (2003). Censored regression quantiles. *J. Amer. Statist. Assoc.* **98,** 1001–1012.

Piantadosi, S. (2005). *Clinical Trials: A Methodologic Perspective*. Wiley series in probability and statistics. Wiley-Interscience, 2nd ed..

Walker, S. and Mallick, B.K., (1999). A Bayesian semiparametric accelerated failure time model. *Biometrics* **55,** 477–483.

Ying, Z., Jung, S. H., Wei, L. J. (1995). Survival analysis with median regression models. *J. Amer. Statist. Assoc.* **90,** 178–184.

Bayesian Analysis of Survival Data with Semi-competing Risks and Treatment Switching

Yuanye Zhang, Qingxia Chen, Ming-Hui Chen, Joseph G. Ibrahim, Donglin Zeng, Zhiying Pan, and Xiaodong Xue

Abstract Treatment switching is common in clinical trials due to ethical and practical reasons. When the outcome of interest is time to death and patients were switched at the time of intermediate nonterminal event, semi-competing risk issue intertwines with the challenge of treatment switching. In this chapter, we develop a Bayesian conditional model for survival data with semi-competing risks in the presence of partial treatment switching. Properties of the conditional model are examined and an efficient Gibbs sampling algorithm is developed to sample from the posterior distribution. A Bayesian procedure to estimate the marginal survival functions and to assess the treatment effect is also derived. The Deviance Information Criterion with an appropriate deviance function and Logarithm of the Pseudo-marginal Likelihood are constructed for model comparison. The proposed method is examined empirically through a simulation study and is further applied to analyze data from a colorectal cancer study.

Y. Zhang (✉)
Novartis Institutes for BioMedical Research, Inc., Cambridge, MA, 02139, USA
e-mail: yuanyevickiezhang@gmail.com

Q. Chen
Department of Biostatistics, Vanderbilt University, Nashville, TN 37232, USA
e-mail: cindy.chen@vanderbilt.edu

M.-H. Chen
Department of Statistics, University of Connecticut, U-4120, Storrs, CT 06269, USA
e-mail: ming-hui.chen@uconn.edu

J.G. Ibrahim • D. Zeng
Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA
e-mail: ibrahim@bios.unc.edu; dzeng@email.unc.edu

Z. Pan • X. Xue
Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320, USA
e-mail: jpan@amgen.com; axue@amgen.com

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_14,

1 Introduction

Treatment switching is common in clinical trials, especially for cancer and other severe diseases. One type of treatment switching is that for patients who are randomized to the control group, they are allowed to switch to the treatment arm after disease progression. Various methods were utilized to analyze the treatment switching data of clinical trials with survival outcomes. The most conventional method is intent-to-treat (ITT) analysis, where the analysis was based on the initial treatment intent, not on the treatment eventually administered. Although ITT analysis can reflect the design of the study, use data of all patients, and avoid some misleading artifacts during intervention, it may underestimate the treatment effect [2]. Other modeling and analysis methods include per-protocol analysis, model with treatment as a time-varying covariate, causal proportional hazards estimator, rank preserving structural failure time models, and parametric randomization-based method. Morden et al. [3] gave a nice review of these methods and conducted simulation studies to compare them under different scenarios. Recently, Zeng et al. [4] proposed a novel class of semi-parametric semi-competing risks survival models to accommodate treatment switching. More references for treatment switching literature are available in [4].

In this chapter, we develop a Bayesian conditional model for survival data with semi-competing risks in the presence of partial treatment switching, that is, not every subject in the control arm switched to experimental treatment. We develop an efficient Gibbs sampling algorithm for the posterior inference of the parameters. Under the Bayesian paradigm, the marginal predictive survival function under the proposed model can be carried out without resorting to asymptotics.

The rest of the chapter is organized as follows. Section 2 presents a detailed development of the semi-competing risks conditional model and provides the explicit expressions for the likelihood function based on the observed data and the Bayesian formulation of the predictive survival function. A simulation study is carried out to examine the empirical performance of the posterior estimates and Bayesian model criteria in Sect. 3. A detailed analysis of a subset of the data from the Panitumumab 408 study is presented in Sect. 4. We conclude the chapter with a brief discussion in Sect. 5.

2 The Semi-competing Risks Conditional Models

2.1 Models

To introduce the proposed model, we use the following notation. As motivated from the Panitumumab 408 study, we consider disease progression as a nonterminating event. However, the proposed model can be applied to any other type of non-terminating event. Let E be a dichotomous variable to denote the disease progression

status of subjects, where $E = 1$ if the subject is in the disease progression population, which include subjects who eventually develop disease progression before death, and $E = 0$ if otherwise. Also let T_D denote the time from study entry to death for subjects with $E = 0$. For the disease progression population ($E = 1$), we further let T_P denote the time from study entry to disease progression and let T_G denote the time from disease progression to death.

The proposed statistical model consists of the following three components. The first component is to model the disease progression status E given the baseline covariates x and the treatment indicator A ($A = 1$ if the subject is on the treatment arm and $A = 0$ if the subject is on the control arm). To this end, we assume

$$\text{logit}\{P(E=1|A,x,\alpha)\} = \log\left\{\frac{P(E=1|A,x,\alpha)}{1-P(E=1|A,x,\alpha)}\right\} = \alpha_0 + A\alpha_1 + x'\alpha_2, \quad (1)$$

where α_0, α_1, and α_2 are unknown coefficients and $\alpha = (\alpha_0, \alpha_1, \alpha_2')'$. The second component models the survival distribution of the non-progression population given x and A, which is defined by

$$h_D(t|A,x,E=0) = h_0(t)\exp\{A\beta_0 + x'\gamma_0\}, \quad (2)$$

where $h_D(t|A,x,E=0)$ is the conditional hazard function of T_D given the covariates, $h_0(t)$ is an unknown baseline hazard function, and (β_0, γ_0) are unknown regression coefficients.

As introduced, T_P and T_G are potentially dependent. To capture this dependence, we assume the following models for T_P and T_G:

$$h_P(t|A,x,E=1) = h_1(t)\exp\{A\beta_1 + x'\gamma_1\} \text{ and} \quad (3)$$

$$h_G(t|A,z,V,T_P,E=1) = h_2(t)\exp\{A\beta_{21} + V(1-A)\beta_{22} + z'\gamma_{21} + T_P\gamma_{22}\}, \quad (4)$$

where $h_P(t|A,x,E{=}1)$ is the conditional hazard function for T_P, $h_G(t|A,z,V,T_P, E{=}1)$ is the conditional hazard function for T_G, both $h_1(t)$ and $h_2(t)$ are unknown baseline hazard functions, and the β's and γ's are regression coefficients. Here, V is the treatment switching indicator (1 if switching; 0 if no switching) and z reflects the covariates collected at baseline or at disease progression, which could be prognostic factors for the treatment switching decision. The model defined by (1)–(4) is called the semi-competing risks conditional model, denoted by CM, and is very flexible to account for the dependence between T_D and T_P. It allows for positive, null, or negative association between T_P and T_G by letting $\gamma_{22} < 0$, $\gamma_{22} = 0$, or $\gamma_{22} > 0$, respectively.

We further assume piecewise constant hazard functions for the baseline hazard functions $h_0(t)$, $h_1(t)$, and $h_2(t)$. For $k = 0,1,2$, let $0 < s_{k1} < s_{k2} < \ldots < s_{kJ_k}$ be a finite partition of the time axis. Thus, we have the J_k intervals: $(0, s_{k1}]$, $(s_{k1}, s_{k2}]$, …, $(s_{k,J_k-1}, s_{kJ_k}]$, where $s_{kJ_k} = \infty$. In the j^{th} interval, we assume a constant baseline

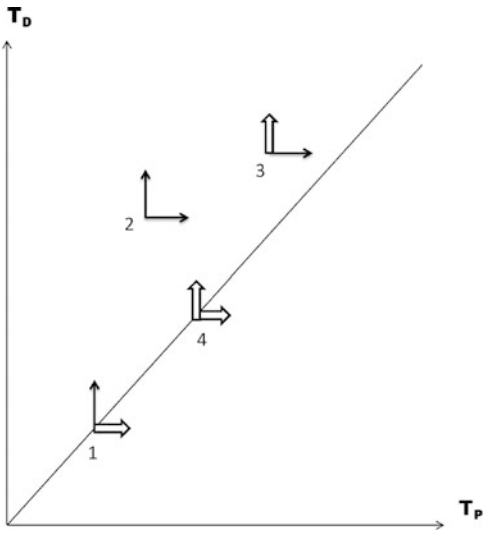


Fig. 1 A closed arrow means observed; an open arrow means censored or equal to ∞. Patients 1, 2, 3, and 4 are examples of categories I, II, III, and IV defined in Sect. 2.2.

hazard, $h_k(y|\lambda_k) = \lambda_{kj}$ for $y \in (s_{k,j-1}, s_{kj}]$. Letting $\lambda_k = (\lambda_{k1}, \lambda_{k2}, \ldots \lambda_{kJ_k})'$, the cumulative baseline hazard function corresponding to $h_k(t)$ is given by

$$H_k(y|\lambda_k) = \lambda_{kj}(y - s_{k,j-1}) + \sum_{g=1}^{j-1} \lambda_{kg}(s_{kg} - s_{k,g-1}), \text{ if } s_{k,j-1} < y \leq s_{kj} \qquad (5)$$

for $k = 0, 1, 2$.

2.2 *Likelihood Function and Posterior Inference*

In addition to the notations defined in Sect. 2.1, we let C denote the independent censoring time, $Y_D = \min(T_D, C)I(E = 0) + \min(T_P + T_G, C)I(E = 1)$, $Y_P = \min(T_P, C)I(E = 1) + \infty I(E = 0)$, $R = I(Y_D < C)$, and $\delta = I(Y_P < C)$, where $I(H) = 1$ if H is true, and 0 otherwise. We let $Y_P = \infty$ when $E = 0$ for mathematical convenience and to emphasize that for the progression-free population with $E = 0$, the progression will never happen. We partition the patient population into four categories: (I) patient died at time Y_D without disease progression and the observed data are $D_I = (E = 0, Y_D, \delta = 0, R = 1, x, A)$; (II) patient was observed to have disease progression at Y_P and then died at Y_D so we have the observation $D_{II} = (E = 1, Y_P, Y_G = Y_D - Y_P, R = 1, \delta = 1, x, A, V(1-A), z)$; (III) patient was observed to have disease progression at Y_P and right censored at Y_D so we have the observation $D_{III} = (E = 1, Y_P, Y_G = Y_D - Y_P, R = 0, \delta = 1, x, A, V(1-A), z)$; and (IV) patient was only observed to be right censored at Y_D and no disease progression occurred before Y_D so we have $D_{IV} = (Y_D, R = 0, \delta = 0, x, A)$. For the patients in the category IV, E is missing with possible values of 0 or 1. The relationships between T_D and T_P from the four categories are illustrated in Fig. 1. As shown in the figure, T_D is greater than

or equal to T_P since when the patient had the terminating event (death), he/she is not at risk of having nonterminating event (disease progression) any more.

Let S_D, S_P, and S_G be the survival functions, and f_D, f_P, and f_G be the density functions, of T_D, T_P, and T_G, respectively, where $S_k(t|\cdot) = \exp(-\int h_k(\mu|\cdot)d\mu)$, $f_k(t|\cdot) = h_k(t|\cdot)S_k(t|\cdot)$, and $h_k(t|\cdot)$ with $k = D,P,G$ are defined in (2), (3), and (4), respectively. Let L_I, L_{II}, L_{III}, and L_{IV} be the likelihood functions of categories I, II, III, and IV, respectively. We have

$$\begin{aligned}
L_I &= \prod_{i \text{ in } (I)} P(E=0|A_i,x_i,\alpha) f_D(y_{Di}|A_i,x_i,\beta_0,\gamma_0,\lambda_0),\\
L_{II} &= \prod_{i \text{ in } (II)} P(E_i=1|A_i,x_i,\alpha) f_P(y_{Pi}|A_i,x_i,\beta_1,\gamma_1,\lambda_1)\\
&\qquad f_G(y_{Gi}|A_i,V_i,z_i,y_{Pi},\beta_2,\gamma_2,\lambda_2),\\
L_{III} &= \prod_{i \text{ in } (III)} P(E_i=1|A_i,x_i,\alpha) f_P(y_{Pi}|A_i,x_i,\beta_1,\gamma_1,\lambda_1)\\
&\qquad S_G(y_{Gi}|A_i,V_i,z_i,y_{Pi},\beta_2,\gamma_2,\lambda_2),\\
L_{IV} &= \prod_{i \text{ in } (IV)} P(E_i=1|A_i,x_i,\alpha) S_P(y_{Di}|A_i,x_i,\beta_1,\gamma_1,\lambda_1)\\
&\qquad + P(E_i=0|A_i,x_i,\alpha) S_D(y_{Di}|A_i,x_i,\beta_0,\gamma_0,\lambda_0). \qquad (6)
\end{aligned}$$

Therefore, the overall likelihood function of the observed data D_{obs} is $L(\theta|D_{obs}) = L_I \times L_{II} \times L_{III} \times L_{IV}$, where $\theta = (\alpha', \beta_0, \gamma_0', \lambda_0', \beta_1, \gamma_1', \lambda_1', \beta_2', \gamma_2', \lambda_2')'$ is the vector of all the model parameters with $\beta_2 = (\beta_{21}, \beta_{22})'$ and $\gamma_2 = (\gamma_{21}', \gamma_{22}')'$.

To conduct Bayesian inference, we need to specify a prior distribution for θ. We assume that α, $(\beta_0, \gamma_0')'$, $(\beta_1, \gamma_1')'$, $(\beta_2', \gamma_2')'$, λ_0, λ_1, and λ_2 are independent and specify the priors for these parameters as: $\alpha \sim N_{p_a}(0, \Sigma_a)$, $(\beta_0, \gamma_0')' \sim N_{p_0}(0, \Sigma_0)$, $(\beta_1, \gamma_1')' \sim N_{p_1}(0, \Sigma_1)$, $(\beta_2', \gamma_2')' \sim N_{p_2}(0, \Sigma_2)$, and $\lambda_{kj} \sim \text{Gamma}(a_{kj}, b_{kj})$ for $j = 1, \ldots, J_k$ and $k = 0, 1, 2$, where p_a, p_0, p_1, and p_2 are the dimensions corresponding to the respective vectors of the model parameters, and Σ_a, Σ_0, Σ_1, Σ_2, a_{kj}, and b_{kj} are pre-specified hyper-parameters. In the simulation study in Sect. 3 and the analysis of the real data from a colorectal cancer study in Sect. 4, these hyper-parameters were specified as $\Sigma_a = 1000 I_{p_a}$, $\Sigma_0 = 1000 I_{p_0}$, $\Sigma_1 = 1000 I_{p_1}$, $\Sigma_2 = 1000 I_{p_2}$, and $a_{kj} = b_{kj} = 0$ for $j = 1, \ldots, J_k$ and $k = 0, 1, 2$, where I_{p_a}, I_{p_0}, I_{p_1}, and I_{p_2} are the identity matrices. We note that when $a_{kj} = b_{kj} = 0$, we specified a Jeffreys prior for each λ_{kj} for $j = 1, \ldots, J_k$ and $k = 0, 1, 2$. Therefore, the resulting joint prior for θ, denoted by $\pi(\theta)$, is essentially noninformative. Using the above prior for θ and the likelihood function $L(\theta|D_{obs})$, the posterior distribution of θ given D_{obs} can be written as

$$\pi(\theta|D_{obs}) \propto L(\theta|D_{obs})\pi(\theta). \qquad (7)$$

To facilitate the posterior computation, we introduce a set of latent variables $E^* = \{E_i^*, i \in (IV)\}$ such that the augmented version of L_{IV} is given by

$$L_{IV}^* = \prod_{i \text{ in } (IV)} [P(E_i = 1|A_i, x_i, \alpha) S_P(y_{Di}|A_i, x_i, \beta_1, \gamma_1, \lambda_1)]^{E_i^*} \left[P(E_i = 0|A_i, x_i, \alpha) S_D(y_{Di}|A_i, x_i, \beta_0, \gamma_0, \lambda_0)\right]^{1-E_i^*}. \quad (8)$$

In order to carry out posterior inference, we develop an efficient Gibbs sampling algorithm to sample θ from the posterior distribution in (7). Let $[A|B]$ denote the conditional distribution of A given B. To run the Gibbs sampling algorithm, we sample from the following conditional distributions in turn: (i) $[\lambda_0, \lambda_1, \lambda_2 | \beta_0, \gamma_0, \beta_1, \gamma_1, \beta_2, \gamma_2, E^*, D_{obs}]$; (ii) $[\beta_0, \gamma_0, \beta_1, \gamma_1, \beta_2, \gamma_2, E^* | \alpha, \lambda_0, \lambda_1, \lambda_2, D_{obs}]$; and (iii) $[\alpha | E^*, D_{obs}]$. For (ii), we use the modified collapsed Gibbs technique [1]. It is easy to show that

$$\begin{aligned} &[\beta_0, \gamma_0, \beta_1, \gamma_1, \beta_2, \gamma_2, E^* | \alpha, \lambda_0, \lambda_1, \lambda_2, D_{obs}] \\ &= [\beta_0, \gamma_0, \beta_1, \gamma_1, \beta_2, \gamma_2 | \alpha, \lambda_0, \lambda_1, \lambda_2, D_{obs}][E^* | \alpha, \beta_0, \lambda_0, \gamma_0, \beta_1, \gamma_1, \lambda_1, D_{obs}]. \end{aligned} \quad (9)$$

Since the conditional distributions in (i), (ii), and (iii) are either standard distributions or the conditional densities are log-concave, the implementation of the Gibbs sampling algorithm is straightforward.

2.3 The Predictive Survival Function with Partial Treatment Switching

An inferential research goal in this research is to compare the survival function of the death time in the setting when no subjects have switched treatment. Let $T_D^*(a)$ denote a potential survival time when a subject receives treatment a at the time of randomization and stays on the same treatment over the entire study duration. Let $S_a(t|\theta) = P(T_D^*(a) > t|\theta)$. Following [4], we state the following two assumptions: (i) Treatment A is completely randomized and $T_D^*(a) = T_D(a)$ if a subject never switches treatment; and (ii) Given $(A = 0, z, T_P = u)$ or $(A = 1, z, T_P = u)$, V is independent of the potential outcomes $\{T_D^*(0), T_D^*(1)\}$. Under Assumptions (i) and (ii), the posterior marginal survival functions $S_a(t|\theta)$ can be estimated by

$$M^{-1} \sum_{m=1}^{M} \int_x \hat{S}_D(t|A = a, x, E = 0, \theta^{(m)}) \hat{P}(E = 0 \mid A = a, x, \theta^{(m)}) \hat{f}_X(x \mid A = a) dx$$

$$+M^{-1} \sum_{m=1}^{M} \int_{x,z,u} \Big[\hat{S}_G(t - u \mid A = a, z, V = 0, T_P, E = 1, \theta^{(m)}) \hat{f}_X(x \mid A = a) \times$$

$$\hat{f}_P(u \mid A = a, x, E = 1, \theta^{(m)}) \hat{f}_Z(z \mid A = a, x, E = 1) \hat{P}(E = 1 \mid A = a, x, \theta^{(m)})\Big] du dz dx,$$

where $\theta^{(m)}$ is the mth MCMC sample from the posterior distribution in (7), $f_X(x|\cdot)$ and $f_Z(z|\cdot)$ are the conditional densities of X and Z, respectively. We estimated $f_X(x|\cdot)$ and $f_Z(z|\cdot)$ nonparametrically as $\hat{f}_X(x \mid A = a) = \sum_{j=1}^{n} I(X_j = x, A_j = a)/\sum_{j=1}^{n} I(A_j = a)$ and

$$\hat{f}_Z(z \mid X, A = a, E = 1) = \frac{\sum_{j\in(II,III)} I(Z_j = z) K_{a_n}(\|X_j - X\|) I(A_j = a)}{\sum_{j\in(II,III)} K_{a_n}(\|X_j - X\|) I(A_j = a)},$$

where K_{a_n} is the kernel function with bandwidth a_n. The survival and density functions of T_D, T_P, and T_G can be estimated by plugining the piecewise exponential model estimates for the baseline hazard functions. In particular, $H_k(y|\lambda_k^{(m)}) = \sum_{j=1}^{J_k}\{I(s_{k,j-1} < t \le s_{kj})[\lambda_{kj}^{(m)}(t - s_{k,j-1}) + \sum_{g=1}^{j-1}\lambda_{kg}^{(m)}(s_{kg} - s_{k,g-1})]\}$ and $h_k(y|\lambda_k^{(m)}) = \sum_{j=1}^{J_k}\{I(s_{k,j-1} < t \le s_{kj})[\lambda_{kj}^{(m)}(t - s_{k,j-1})I(j = 1) + \lambda_{k,j-1}^{(m)}(s_{kj} - s_{k,j-1})I(j > 1)]\}$ with $s_{k0} = 0$ for $k = 0, 1, 2$.

3 A Simulation Study

To examine the empirical performance of the posterior estimates as well as Deviance Information Criterion (DIC) and Logarithm of the Pseudo-Marginal Likelihood (LPML), we carry out a simulation study. Five hundred simulated datasets with $n = 500$ as well as $n = 1{,}000$ were generated. In the simulation study, the baseline treatment A was generated from a Bernoulli(0.5), corresponding to a randomized trial with a 1 : 1 sample size allocation; two baseline covariates X_1 and X_2 were independently generated from a $U(-1, 1)$ and a Bernoulli(0.6), respectively. Given A and (X_1, X_2), E was generated from model (1) with the coefficients (including an intercept) being 1.6, -1.8, 1, and 0.1, respectively. When $E = 0$, we simulated T_D from model (2) with $H_0(t) = t$, $\beta_0 = -1$ and $(\gamma_{01}, \gamma_{02}) = (1, 0.2)$. For $E = 1$, T_P was generated with $H_1(t) = 5t$, $\beta_1 = -0.5$ and $(\gamma_{11}, \gamma_{12}) = (1, 0)$ and an additional prognostic factor Z at disease progression was generated from a $U(0, 10)$ while the selection into treatment switching (V) for a subject in the control arm ($A = 0$) was from a Bernoulli(p), where $p = \exp(-0.5 + 0.3T_P + 0.2X_1 + 0.5Z)/[1 + \exp(-0.5 + 0.3T_P + 0.2X_1 + 0.5Z)]$. Moreover, T_G was generated with $H_2(t) = t, \beta_{21} = -0.3, \beta_{22} = -0.5$, $\gamma_{21} = -0.5, \gamma_{22} = 0.5, \gamma_{23} = -0.4$ and an additional parameter $\gamma_{24} = 0.6$ as the coefficient for T_P. Finally, the censoring time was generated from a $U(1, 7)$ and the study duration was $\tau^* = 3$. The latter yielded the average proportions of categories I to IV as 23 %, 55 %, 10 %, 12 %.

For each simulated dataset of $n = 1{,}000$, we fit the proposed CM with various values of (J_0, J_1, J_2) and computed DIC and LPML. The mean values of the DICs and LPMLs over the 500 simulated datasets were 2247.50 and -1123.92 for $(J_0, J_1, J_2) = (1, 1, 1)$; 2259.12 and -1129.84 for $(J_0, J_1, J_2) = (5, 5, 5)$; and 2274.35 and -1137.61 for $(J_0, J_1, J_2) = (10, 10, 10)$. We note that the true value of (J_0, J_1, J_2) is $(1, 1, 1)$. Thus, both DIC and LPML correctly identified the true model. Under the

Table 1 Posterior estimates under CM in the simulation study

		$n = 500$					$n = 1,000$				
Parameter	True	EST	SD	SE	RMSE	CP%	EST	SD	SE	RMSE	CP%
					T_D model						
β_0	−1.00	−1.01	0.21	0.22	0.22	0.94	−1.01	0.14	0.15	0.15	0.94
γ_{01}	1.00	1.01	0.17	0.18	0.18	0.95	1.01	0.12	0.13	0.13	0.94
γ_{02}	0.20	0.21	0.19	0.20	0.20	0.95	0.21	0.14	0.14	0.14	0.94
					T_P model						
β_1	−0.50	−0.51	0.12	0.12	0.12	0.94	−0.51	0.09	0.08	0.08	0.95
γ_{11}	1.00	1.01	0.11	0.11	0.11	0.94	1.00	0.07	0.07	0.07	0.95
γ_{12}	0.00	0.00	0.12	0.13	0.13	0.93	0.00	0.08	0.09	0.09	0.94
					T_G model						
β_{21}	−0.30	−0.30	0.15	0.15	0.15	0.95	−0.30	0.10	0.10	0.10	0.95
β_{22}	−0.50	−0.50	0.15	0.16	0.16	0.94	−0.51	0.11	0.11	0.11	0.93
γ_{21}	−0.50	−0.51	0.12	0.12	0.12	0.94	−0.50	0.08	0.09	0.09	0.94
γ_{22}	0.50	0.50	0.13	0.13	0.13	0.94	0.50	0.09	0.09	0.09	0.94
γ_{23}	−0.40	−0.41	0.21	0.22	0.22	0.95	−0.40	0.15	0.14	0.14	0.97
γ_{24}	0.60	0.57	0.22	0.21	0.22	0.96	0.59	0.15	0.15	0.15	0.95
					E model						
α_0	1.60	1.63	0.22	0.23	0.23	0.93	1.62	0.15	0.16	0.16	0.94
α_1	−1.80	−1.83	0.23	0.23	0.24	0.94	−1.82	0.16	0.16	0.16	0.95
α_{21}	1.00	1.02	0.19	0.19	0.19	0.96	1.02	0.14	0.13	0.13	0.96
α_{22}	0.10	0.10	0.22	0.21	0.21	0.96	0.11	0.15	0.15	0.15	0.96
					Estimated Survival function of control arm						
$S_0(\tau^*/2)$	0.25	0.25	0.02	0.02	0.02	0.94	0.25	0.03	0.03	0.03	0.92
$S_0(\tau^*)$	0.07	0.07	0.01	0.01	0.01	0.93	0.07	0.01	0.02	0.02	0.93
					Estimated Survival function of treatment arm						
$S_1(\tau^*/2)$	0.48	0.48	0.02	0.02	0.02	0.95	0.48	0.02	0.03	0.03	0.94
$S_1(\tau^*)$	0.24	0.24	0.02	0.02	0.02	0.93	0.25	0.02	0.02	0.03	0.92

best combination of (J_0, J_1, J_2), namely, $(1,1,1)$, the average of the posterior means (EST), and the average of the posterior standard deviations (SD), the simulation standard error (SE), the root of the mean squared error (RMSE), and the coverage probability (CP) of the 95 % highest posterior density (HPD) intervals for each parameter as well as $S_a(t|\theta)$ were computed. The results are given in Table 1. Table 1 shows excellent empirical performance of the posterior estimates for all the parameters as well as the survival probabilities for both $n = 500$ and $n = 1,000$. In particular, the ESTs are nearly identical to the true values, the SDs are very close to the SEs, and the CPs are very close to 95 %.

4 Application to a Colorectal Cancer Study

The proposed CM model is applied to analyze the data collected from a colorectal cancer study, the panitumumab study conducted by Amgen Inc. Panitumumab study was an open label, randomized, phase III multicenter study designed to compare the efficacy and safety of panitumumab plus best supportive care (P+BSC) versus BSC alone in patients with EGFr-expressing metastatic colorectal cancer. Patients were randomly assigned to receive P+BSC (A=1) or BSC (A=0). The baseline covariates, x, include age in years at screening (Age), baseline Eastern Cooperative Oncology Group score (score 0 or 1 versus ≥ 2 (bECOG01)), primary tumor diagnosis type (Rectal), gender (Male), and region (western Europe (WesternEU), eastern and central Europe (CenEstEU), versus rest of the world). The covariates, z, in the T_G model include bECOG01, age at disease progression (PR Age), best tumor response with partial response (BTR PR) or stable disease (BTR SD) versus progressive disease, last ECOG score (LECOG01), and grade 2 or above adverse events (AE).

We fit CM with different values of J_0, J_1, and J_2 to the panitumumab data. The DIC and LPML values of the seven combinations of (J_0, J_1, J_2) are given in Table 2. The model based on $(J_0, J_1, J_2) = (1, 30, 5)$ achieves the smallest DIC value and the largest LPML value. Table 3 gives the posterior estimates of the CM model with $(J_0, J_1, J_2) = (1, 30, 5)$. The 95 % HPD intervals for treatment were (-1.754, -0.486) under the E model, (-1.155, 0.156) under the T_D model, (-1.388, -0.936) under the T_P model, and (-1.196, -0.362) under the T_G model. These results imply that treatment is associated with E, T_P, T_G but not with T_D. Comparing to the BSC patients who had progression but didn't switch to P+BSC group, the patients switching the treatment group had significantly lower hazard of death with posterior mean and 95 % HPD interval of -1.371 and (-1.773, -0.971). The posterior mean and 95 % HPD interval for γ_{22} in (4) were -0.083 and (-0.161, -0.008), which implies that there is a positive association between T_P and T_G.

In all of the Bayesian computations, we used 20,000 Gibbs samples after 1,000 burn-ins for each model to compute all the posterior estimates, including posterior means, posterior standard deviations, and 95 % HPD intervals. Codes were written for the FORTRAN 95 compiler using IMSL subroutines with double precision accuracy. The convergence of the Gibbs sampler was checked using several diagnostic procedures discussed in [1]. The autocorrelations for all model parameters disappeared after lag 10.

5 Discussions

In this chapter, we have developed a Bayesian conditional model for partial treatment switching problems with terminal event as primary outcome and the non-terminal event time as switching time. The proposed model accommodates treatment switching and flexible dependence between the progression time and

Table 2 DIC, p_D, and LPML values for the Panitumumab data

Parameter			CM		
J_0	J_1	J_2	DIC	p_D	LPML
1	30	5	3482.64	67.32	−1746.76
3	30	5	3486.14	69.48	−1748.76
5	30	5	3486.92	71.44	−1748.94
1	25	5	3500.00	62.20	−1755.11
1	35	5	3484.61	72.54	−1748.62
1	30	3	3488.20	65.37	−1749.05
1	30	10	3490.95	72.46	−1751.73

Table 3 Posterior estimates for the Panitumumab data under CM with $(J_0, J_1, J_2) = (1, 30, 5)$

Parameter	EST	SD	95 % HPD	Parameter	EST	SD	95% HPD
			E Model				T_D Model
Intercept	1.475	0.983	(−0.474, 3.369)				
Treatment	−1.114	0.326	(−1.754, −0.486)	Treatment	−0.488	0.334	(−1.155, 0.156)
Age	−0.010	0.014	(−0.038, 0.017)	Age	0.024	0.015	(−0.004, 0.054)
bECOG01	1.992	0.340	(1.309, 2.641)	bECOG01	−0.617	0.296	(−1.180, −0.021)
Rectal	0.308	0.334	(−0.346, 0.950)	Rectal	−0.058	0.317	(−0.667, 0.573)
Male	−0.320	0.329	(−0.966, 0.320)	Male	−0.329	0.307	(−0.933, 0.275)
CenEastEU	−0.003	0.615	(−1.186, 1.215)	CenEastEU	−0.312	0.631	(−1.534, 0.929)
WesternEU	0.342	0.418	(−0.461, 1.176)	WesternEU	0.239	0.401	(−0.540, 1.021)
			T_P Model				T_G Model
Treatment	−1.159	0.115	(−1.388, −0.936)	Treatment	−0.783	0.214	(−1.196, −0.362)
Age	−0.015	0.005	(−0.025, −0.006)	V*(1-Treatment)	−1.371	0.208	(−1.773, −0.971)
bECOG01	−0.777	0.176	(−1.110, −0.429)	PR Age	−0.004	0.005	(−0.014, 0.006)
Rectal	−0.036	0.109	(−0.246, 0.179)	BTR PR	−0.243	0.347	(−0.920, 0.434)
Male	−0.041	0.109	(−0.252, 0.174)	BTR SD	−0.166	0.173	(−0.502, 0.177)
CenEastEU	0.226	0.233	(−0.228, 0.681)	bECOG01	−0.244	0.196	(−0.619, 0.141)
WesternEU	−0.049	0.159	(−0.367, 0.255)	LECOG01	−1.033	0.145	(−1.320, −0.749)
				AE	0.290	0.116	(0.064, 0.516)
				Prog Time	−0.083	0.039	(−0.161, −0.008)

survival time. This type of scenarios arise often in clinical trials, in which patients immediately switch to the experimental treatment once he/she experiences an event (for example, progression). The proposed model is the parallel Bayesian development of the approach developed in [4]. The Bayesian estimates were very similar to those based on the frequentist method. For example, for the proposed Bayesian approach, the treatment effect estimates (standard deviations) in the E, T_D, T_P, and T_G models were -1.114 (0.326), -0.488 (0.334), -1.159 (0.115), and -0.783 (0.214), respectively. For the frequentist method mentioned in [4], the corresponding estimates (standard errors) were -1.070 (0.319), -0.464 (0.347), -1.144 (0.118), and -0.784 (0.214). Thus, these two sets of the estimates were very similar. However, under the Bayesian paradigm, the marginal predictive survival function under the

proposed model can be carried out straightforward without resorting to asymptotics. The proposed model is useful for researchers encountering treatment switching studies in the presence of semi-competing risks where one is interested in assessing the treatment effect on survival of terminal event.

References

[1] M.-H. Chen, Q.M. Shao, and J.G. Ibrahim, *Monte Carlo Methods in Bayesian Computation*, (2000).

[2] S. M. Marcus and R. D. Gibbons, *Estimating the efficacy of receiving treatment in randomized clinical trials with noncompliance*, Health Services & Outcomes Research Methodology **2** (2001), 247–58.

[3] James P Morden, Paul C Lambert, Nicholas Latimer, Keith R Abrams, and Allan J Wailoo, *Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study*, BMC Medical Research Methodology **11** (2011), no. 4, 1–20.

[4] D. Zeng, Q. Chen, M.-H. Chen, J.G. Ibrahim, and Amgen research group, *Estimating treatment effects with treatment switching via semi-competing risks models: An application to a colorectal cancer study*, Biometrika **99** (2012), 167–184.

Part IV
High Dimensional Data Analysis and Statistical Genetics

High-Dimensional Ordinary Differential Equation Models for Reconstructing Genome-Wide Dynamic Regulatory Networks

Shuang Wu, Zhi-Ping Liu, Xing Qiu, and Hulin Wu

Abstract The gene regulatory network (GRN) is a complex control system and plays a fundamental role in the physiological and development processes of living cells. Focusing on the ordinary differential equation (ODE) modeling approach, we propose a novel pipeline for constructing high-dimensional dynamic GRNs from genome-wide time course gene expression data. A five-step procedure, i.e., detection of temporally differentially expressed genes, clustering genes into functional modules, identification of network structure, parameter estimate refinement and functional enrichment analysis, is developed, combining a series of cutting-edge statistical techniques to efficiently reduce the dimension of the problem and to account for the correlations between measurements from the same gene. In the key step of identifying the network structure, we employ the advanced parameter estimation and statistical inference methods to perform model selection for the ODE models. The proposed pipeline is a computationally efficient data-driven tool bridging the experimental data and the mathematical modeling and statistical analysis. The application of the pipeline to the time course gene expression data from influenza-infected mouse lungs has led to some interesting findings of the immune process in mice and also illustrated the usefulness of the proposed methods.

1 Introduction

The genome encodes thousands of genes whose products enable cell survival and numerous cellular functions, such as cell growth and division, the response to environmental stimuli, and so on. The interactions of these genes and their products

S. Wu • Z.-P. Liu • X. Qiu • H. Wu (✉)
Department of Biostatistics and Computational Biology, University of Rochester,
601 Elmwood Avenue, Rochester, NY 14642, USA
e-mail: shuang_wu@urmc.rochester.edu; zhiping_liu@urmc.rochester.edu;
xing_qiu@urmc.rochester.edu; hulin_wu@urmc.rochester.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_15,

form intricate regulatory networks. Understanding the dynamics of these networks sheds light on the mechanisms of diseases that occur when certain regulations are dysfunctional. With the rapid advances in high-throughput technologies such as DNA microarray and next generation RNA-seq, tremendous amounts of gene expression data have been produced, especially time course gene expression data that allow investigators to study dynamic behaviors of the genes. Analyzing these data with statistical and computational approaches can provide quantitative measures of the global response of a biological system to specific interventions and help us better understand the biological process in a systematic way.

The abundance of genomic data has largely promoted the development of methodologies for data-driven inference of gene regulatory networks (GRNs). The objective is to reconstruct the GRNs from experimental data using mathematical models, especially dynamic network models that aim to capture the complex phenomena of biological systems by modeling the time course gene expression data. Many models have been proposed, including the information theory models (Steuer et al. 2002; Stuart et al. 2003), Boolean networks (Shmulevich et al. 2002; Thomas 1973), vector autoregressive (VAR) and state space models (SSM) (Hirose et al. 2008; Kojima et al. 2009), Bayesian networks (Friedman et al. 2000; Heckerman 1996; Perrin et al. 2003; Zou and Conzen 2005), and differential equation models (Chen et al. 1999; Holter et al. 2001; Lu et al. 2011; Yeung et al. 2002). Excellent reviews on diverse data-driven modeling schemes and related topics can be found in De Jong (2002) and Hecker et al. (2009).

In this paper we focus on the ordinary differential equation (ODE) modeling approach for reconstructing the high-dimensional GRN. In ODE network models, gene regulations are modeled by rate equations, which quantify the change rate (derivative) of the expression of one gene in the system as a function of expression levels of all related genes. It is a directed network graph model and the dynamic feature of the GRN is automatically and naturally quantified. Both up and down regulations as well as self-regulations can be appropriately captured by the ODE model. The general form of the ODE model can be written as

$$X'(t) = F(t, X(t), \theta), \tag{1}$$

where $X(t) = (X_1(t), \ldots, X_G(t))^T$ is a vector representing the gene expression levels of genes $1, \ldots, G$ at time t, $t \in [t_0, T]$, $0 \le t_0 < T < \infty$, and $X'(t)$ is the first order derivative of $X(t)$. The link function F describes the regulatory effects between genes, which possibly depends on a vector of parameters θ.

The identification of the parameters in (1) requires some constraints on the model structure, for instance, specification of the form of the function F. Many previous works have assumed that F is linear, due to the simplicity of linear models (Bansal et al. 2006; Holter et al. 2001; Yeung et al. 2002). A linear ODE model can be written as

$$X_g'(t) = \alpha_{g0} + \sum_{i=1}^{G} \alpha_{gi} X_i(t), \quad g = 1, \ldots, G, \tag{2}$$

where α_{g0} is the intercept and coefficients $\alpha = \{\alpha_{gi}\}_{g,i=1,\ldots,G}$ quantify the regulation effects of other genes on the rate of expression change of the g-th gene. For a small-scale ODE-based GRN model (i.e., G is small), some standard methods such as the least squares method or likelihood-based method can be used to perform statistical inference for the dynamic parameters α. However, when G becomes large, model (2) includes many parameters and the standard statistical methods may fail due to the curse of dimensionality.

The methodologies presented in this paper are useful data-driven tools for reconstructing large-scale GRNs from genome-wide time course gene expression data by ODE models. In a typical gene expression experiment, tens of thousands of genes are measured simultaneously, but only a fraction of them are associated with the biological process of interest or a particular stimulus, such as a therapeutic treatment or virus infection. Since it is reasonable to include only these "responsive" genes in the ODE network model, the first step in the GRN modeling is to identify temporally differentially expressed genes, i.e., genes with expressions changed significantly over time. Within the set of differentially expressed genes, which usually ranges from several hundreds to thousands, many genes behave similarly during the experimental period, making it difficult to distinguish their expression patterns based on the time course data. We propose to cluster these similarly behaved genes (co-expressed genes) into functional modules (Luan and Li 2003; Ma et al. 2006). By doing this, the dimension of the ODE model (2) can be significantly reduced and we can rewrite (2) for functional modules as

$$M_k'(t) = \beta_{k0} + \sum_{i=1}^{K} \beta_{ki} M_i(t), \quad k = 1, \ldots, K, \tag{3}$$

where K is the number of functional modules and $M_k(t)$ is the mean expression curve of the k-th module. The between-gene variation within each module is accounted using mixed-effects modeling approaches. These K functional modules are the nodes of the GRN and when β_{ki} is nonzero, there is a directed edge between module i and module k. Since biological systems are seldom fully connected and most nodes are only directly connected to a small number of other nodes (Jeong et al. 2000), it is commonly assumed that the GRN is a sparse network, i.e., the number of nonzero β_{ki} in model (3) is small. Therefore, to identify the significant edges $\mathcal{S} = \{1 \le k, i \le K : \beta_{ki} \ne 0\}$ from the time course experimental data is an essential step in the reconstruction of a GRN.

Although the model size K of (3) is comparably smaller than that of the original model (2), simultaneous model selection and parameter optimization of (3) are still computationally very expensive, because these involve costly numerical integration and complicated parameter regularization. To tackle this difficulty, we propose to adopt the two-stage smoothing-based estimation method (Liang and Wu 2008; Voit and Almeida 2004), which decouples the system of differential equations into a set of pseudo-regression models by substituting the differentials and state variables in (3) with the estimated derivatives and expression curves from the observed

data. This method avoids numerically solving the differential equations directly and does not need the initial or boundary conditions of the state variables. More importantly, it allows us to perform model selection and parameter estimation for one equation at a time, which significantly reduces the computational cost. These favorable properties of the two-stage smoothing-based estimation method greatly outweigh its disadvantage of a small loss of estimation accuracy in dealing with high-dimensional ODE models (Lu et al. 2011).

In this paper, we propose a novel pipeline to identify dynamic GRNs from genome-wide time course gene expression data by combining a series of cutting-edge statistical techniques. The remainder of the paper is organized as follows. The methodologies in the pipeline are presented in Sect. 2: we first introduce methods to select temporally differentially expressed genes from the genome-wide time course data in Sect. 2.1; in Sect. 2.2, we discuss various methods to cluster these differential genes into functional modules; Sect. 2.3 provides details of the statistical methods for the network structure identification based on the linear ODE model; Sect. 2.4 deals with the parameter refinement of the reconstructed GRN in Sect. 2.3; Sect. 2.5 addresses the functional enrichment analysis of the functional modules and the reconstructed GRN. We then illustrate the proposed pipeline with the time course gene expression data from influenza-infected mouse lungs (Pommerenke et al. 2012) in Sect. 3. We summarize the pipeline with some remarks and discuss perspectives for future research in Sect. 4.

2 Methodology

2.1 Identifying Temporally Differentially Expressed Genes

We adopt the significant testing method proposed by Wu and Wu (2013) to identify temporally differentially expressed genes. It is assumed that the expression profile of each gene $X_g(t)$ is a smooth function of time and the time course measurements are collected as discrete observations from $X_g(t)$ that are contaminated by noisy signals. Denote $U_{gjr} = X_g(t_j) + \epsilon_{gjr}$ as the r-th replicate of the expression measurements of the g-th gene at time t_j, $g = 1,\ldots,G$, $j = 1,\ldots,n$, $r = 1,\ldots,R_j$, where n is the number of sampling time points for each gene and R_j is the number of replicates at t_j. The noises ϵ_{gjr} are assumed to be i.i.d r.v.'s with mean 0 and variance σ^2. A key step is to create an estimate of $X_g(t)$ from the noisy time course data through a data-based eigen-representation: $X_g(t) \approx \mu_g + \sum_{l=1}^{L} \xi_{gl}\phi_l(t)$, where μ_g is the mean expression, ϕ_l are the sequences of orthonormal eigenfunctions and ξ_{gl} are the corresponding functional principal component scores. The top L eigenfunctions are selected such that the total variation explained exceeds a pre-specified threshold (such as 90 %).

In the case of a single experimental group, we are often interested in discovering genes whose expression profiles are time-dependent, where the hypothesis can be written as

$$\mathrm{H}_{g0}: X_g(t) = \mu_g, \text{ v.s. } \mathrm{H}_{g1}: X_g(t) \neq \mu_g, \quad \text{for any } t \in [t_0, T]. \tag{4}$$

The test statistic is a modified F-statistic, which compares the goodness-of-fit of the null model to the alternative model:

$$F_g = \frac{\mathrm{RSS}_g^0 - \mathrm{RSS}_g^1}{\mathrm{RSS}_g^1 + \rho}, \tag{5}$$

where RSS_g^0 and RSS_g^1 are the residual sum of squares under the null and the alternative models for the g-th gene, respectively. This statistic can also be viewed as the signal-to-noise ratio of each gene. For genes with low signal levels, variances in F_g can be high because of small values of RSS_g^1. The small constant ρ in the denominator can help stabilize the variance of F_g and a suggested choice by Wu and Wu (2013) is $\rho = \hat{\sigma}^2$, the estimated variance of the noisy signal. A permutation test is used to generate the null distribution of $(F_1, \ldots, F_G)$ and the multiple testing adjustment method proposed by Benjamini and Hochberg (1995) is applied to control the false discovery rate (FDR). If multiple experimental groups are involved, we want to identify genes with different expression profiles in different groups. See Wu and Wu (2013) for testing across multiple groups and more technical details.

2.2 *Clustering Genes into Functional Modules*

For each differentially expressed gene, we first combine the replicated gene expressions by taking their average at each time point. This is mainly for technical convenience, since most of the existing clustering algorithms do not consider multiple replicates in the time course expression data. We do not recommend averaging the replicates before the significance test because some genes may have outliers in the observed data and averaging replicates for these genes may distort their expression patterns and reduce the testing power. The significance test selects genes with large signal-to-noise ratios, which less likely include any outliers. Next, a gene-wise standardization procedure is applied. It has been well documented that genes with similar biological functions usually have similar expression patterns but with different expression magnitudes (fold changes) (Pommerenke et al. 2012). This standardization procedure can remove such magnitude differences and is commonly adopted when the aim is to produce groups of genes that are functionally related (Eisen et al. 1998).

Many clustering methods have been applied to time course gene expression data, including K-means clustering (Hartigan and Wong 1979), self-organizing maps (SOM) (Kohonen 1997), and hierarchical clustering (Eisen et al. 1998). To account for the dependency of expression levels across time for a given gene, Ma et al. (2006) proposed a nonparametric mixed-effects smoothing spline method under the framework of a mixture distribution to cluster time course gene expression

data. The maximum penalized likelihood method coupled with a variation of the EM algorithm was employed to estimate the model parameters and the number of clusters was determined by the BIC criterion. In practice, we find that this method is computationally very slow. Moreover, the BIC criterion often keeps decreasing as the number of clusters increases, failing to select the number of clusters. For the data application in Sect. 3, we apply K-means clustering to group the differentially expressed genes into functional modules. The most appealing advantage of K-means clustering is its computational efficiency. In addition, we find that for a given number of clusters, the result by K-means clustering is very close to that produced by Ma et al. (2006), since the latter uses K-means to initiate the program. Therefore, we recommend the use of K-means clustering for practical applications if one has limited computational resources.

We propose an empirical rule to determine the number of clusters when using K-means clustering. Denote the total within-cluster sum of squares of the k-th cluster by WCSS_k. We plot the relative changes of the within-cluster sum of squares $(\mathrm{WCSS}_k - \mathrm{WCSS}_{k+1})/\mathrm{WCSS}_k$ against the number of clusters k and choose the number K at the knee of the curve. The motivation behind this proposal is that the within-cluster variation WCSS_k goes down as k increases and the decreasing rate should significantly slows down after k passes the optimal number of clusters. To facilitate the ODE network modeling and functional enrichment analysis in the following steps, we recommend that the number of clusters K should not be chosen too small and ideally it is more than 10.

2.3 Identification of Network Structure

In this section, we present methods to identify the significant regulations between functional modules in model (3). We first obtain the estimates of $M_k(t)$ and their derivatives $M_k'(t)$ from the observed data. Substituting these estimates into (3) can decouple the K-dimensional ODEs into K independent pseudo-regression models. So the problem of identifying the ODE model structure can be transformed into variable selection problems of these K regression models. The decoupling property is a unique feature of the two-stage smoothing-based estimation method for ODEs (Liang and Wu 2008; Voit and Almeida 2004). The benefits of this approach are that it avoids numerically solving the differential equations and allows independent model selection and parameter estimation for one equation at a time, which significantly reduce the computational cost (Lu et al. 2011).

2.3.1 Nonparametric Smoothing

With a slight abuse of notation, we use U_{gj} to denote the expression level at t_j for the g-th gene after combining the replicates and the standardization procedure. Within each of the K functional modules, the gene expression patterns are similar, so we can

treat the time course data of these genes as longitudinal measurements and model them using the following nonparametric mixed-effects model:

$$U_{gj} = M_k(t_j) + \eta_{gi}(t_j) + \varepsilon_{gij}, \quad g \in C_k, \tag{6}$$

where C_k is the collection of gene indices for the k-th module, $M_k(t)$ is the mean expression curve for module k, and $\eta_{gi}(\cdot)$ is the random-effects function that quantifies the deviation of the expression level of gene g from the mean expression $M_k(t)$.

There is a sizable literature about the model fitting of (6) and the methods include smoothing spline, regression spline, penalized spline, and local polynomial smoothing. We omit the detailed presentations of these standard smoothing methods and refer to Wu and Zhang (2005) and references therein for a good review on this topic. Applying any of these nonparametric smoothing techniques, we can obtain the estimates of the mean expression curve $M_k(t)$ and its first order derivative $M_k'(t)$ for each functional module. Following Liang and Wu (2008), we suggest undersmoothing these curve estimates in this step.

2.3.2 Variable Selection for the Linear ODE Model

We plug the estimated mean expression curves and their derivatives $\hat{M}_k(t)$ and $\hat{M}_k'(t)$ into the ODE system (3) to form a set of pseudo-linear regression models. Since $\hat{M}_k(t)$ and $\hat{M}_k'(t)$ are estimated continuously as nonparametric functions, we recommend using augmented data from $\hat{M}_k(t)$ and $\hat{M}_k'(t)$ at time points $t_1^*, \ldots, t_N^*$ ($\in [t_0, T]$), where N can be larger than the original sample size n. This data augmentation strategy has also been used by other investigators before (Bansal et al. 2006) and has been shown to help better estimate the ODE parameters. Denote the augmented data as $y_{kj} = \hat{M}_k'(t_j^*)$ and $z_{kj} = \hat{M}_k(t_j^*)$, $j = 1, \ldots, N$. We can write the pseudo-regression models as

$$y_{kj} = \beta_{k0} + \sum_{i=1}^{K} \beta_{ki} z_{ij} + \delta_{kj}, \quad k = 1, \ldots, K. \tag{7}$$

The error term δ_{kl} represents the aggregated estimation error of $\hat{M}_k(t)$ and $\hat{M}_k'(t)$ and model error due to the substitution of the differential equation variables by $\hat{M}_k(t)$ and $\hat{M}_k'(t)$. Note that these errors are dependent, and the covariates z_{ij} and responses y_{kj} in (7) are derived from the smoothing estimates rather than directly measured data. Therefore, model (7) is not a standard regression model and this is why we refer to it as a "pseudo" linear regression model.

For linear regression models, many penalized methods have been proposed in the regularization framework to conduct variable selection and estimation, such as the least absolute shrinkage and selection operator (Tibshirani 1996) and smoothly clipped absolute deviation (SCAD) (Fan and Li 2001). Many efforts have been made to demonstrate the statistical properties of these methods in high-dimensional settings, such as Zhang and Huang (2008) and Kim et al. (2008) among others. These methods can be applied to perform variable selection for model (7), but their

theoretical properties under the pseudo-linear regression setting need to be carefully considered. In this paper, we focus on the SCAD method for model (7).

Without loss of generality, we assume that both the response $y_k = (y_{k1}, \ldots, y_{kN})^T$ and covariates $z_i = (z_{i1}, \ldots, z_{iN})^T$, $i = 1, \ldots, K$ in (7) are centered, so $\beta_{k0} = 0$. Consider the following penalized objective function

$$\frac{1}{2N} \sum_{j=1}^{N} (y_{kj} - z_i^T \beta_k)^2 + \sum_{i=1}^{K} J_{\lambda_k}(|\beta_{ki}|), \tag{8}$$

where $\beta_k = (\beta_{k1}, \ldots, \beta_{kK})^T$ and $J_\lambda(|\beta|)$ is the SCAD penalty (Fan and Li 2001). We employ the CCCP-SCAD algorithm developed by Kim et al. (2008) to minimize (8). This algorithm is shown to be less sensitive to the initial value, faster, more stable than previous algorithms for the SCAD problem and more importantly, it is always guaranteed to converge to a local minimum (Kim et al. 2008; Lu et al. 2011). Applying the CCCP-SCAD algorithm to each of the K pseudo-linear regression models, we obtain the collective set of nonzero coefficients $\hat{\mathcal{S}} = \{(k,i) : 1 \le k, i \le K, \hat{\beta}_{ki}^S \ne 0\}$, where $\hat{\beta}_{ki}^S$ are the minimizers of (8). This set $\hat{\mathcal{S}}$ gives the structure of the ODE network model (3) and the regulation relationships between the functional modules in the network. Lu et al. (2011) have proved that the SCAD estimator $\hat{\beta}_{ki}^S$ has an oracle property, which ensures the correct recovery of the true network asymptotically.

2.4 *Refining Parameter Estimates*

In Sect. 2.3, the two-stage smoothing-based estimation method is employed to simplify the computation of the ODE models and also facilitate the variable selection procedure. However, the parameter estimates from the two-stage method are not efficient in terms of estimation accuracy, because of the approximation errors brought in by the estimates of the mean expression curves $M_k(t)$ of the functional modules and their derivatives $M_k'(t)$. These errors could be quite large when the data are measured at a sparse grid or with large noise signals. To overcome the estimation deficiency of the two-stage method, we propose to refine the parameter estimates for the selected ODE model using the nonlinear least squares (NLS) method. The parameter estimates from the two-stage method in Sect. 2.3 can be used as the initial estimates in the NLS procedure.

Recall that model (3) is based on the mean expression values of the functional modules. Therefore, the refined parameter estimate obtained from the NLS method provides a description of the average regulatory effect that the genes in one module have on another module. To obtain the gene-specific regulatory parameter estimates, we consider the following mixed-effects ODE model for module k ($1 \le k \le K$):

$$\frac{dX_g(t)}{dt} = \beta_{k0} + \sum_{i \in \hat{\mathcal{S}}_k} \alpha_{gi} M_i(t) = \beta_{k0} + \sum_{i \in \hat{\mathcal{S}}_k} (\beta_{ki} + \gamma_{gi}) M_i(t), \quad g \in C_k, \tag{9}$$

where $X_g(t)$ is the true expression curve of the g-th gene; the set C_k includes the gene indices in module k and $\hat{\mathcal{S}}_k = \{i : 1 \le i \le K, \hat{\beta}_{ki}^S \ne 0\}$ is the collection of nonzero coefficients for the k-th differential equation; the random effects γ_{gi} are assumed to follow normal distributions and they characterize the between-gene variation in the k-th module. Model (9) indicates that gene g in the k-th functional module is regulated by the mean effects of genes in the other modules and the gene-specific effect α_{gi} can be considered as the average regulatory effect β_{ki} with a random deviation γ_{gi}. Lu et al. (2011) recommended using the stochastic approximation EM (SAEM) algorithm (Delyon et al. 1999; Kuhn and Lavielle 2004) to obtain the maximum likelihood estimates (MLEs) of α_{gi}. When the conditional distribution of random effects γ_{gi} does not have a closed-form expression, a Markov chain Monte Carlo (MCMC) method can be used. See Delyon et al. (1999) and Kuhn and Lavielle (2004) for more details about the SAEM algorithm.

2.5 Functional Enrichment Analysis

Up to this point, we have constructed the module-based GRN through the ODE network models using data-driven methods. It is crucial to interpret these data-driven results and summarize the biological implications of the reconstructed GRN. To this end, we perform the functional enrichment analysis for the functional modules and the reconstructed GRN.

Genes within the same functional module may have many biological functions and certain functions may be over-represented in this module compared to the population of genes in an organism or a biological process. These over-represented functions or enriched functions are the key factors to understand the role that the functional module plays in the whole network. We can use DAVID (Huang et al. 2009) to identify the gene ontology (GO) functional annotations and KEGG/BioCarta/Reactome pathways that are enriched in each functional module. In this analysis, a modified Fisher's exact test is carried out for each functional term under the null hypothesis that this function is not over-represented in the functional module compared to the background population (Huang et al. 2009). The statistical significance of each functional term is adjusted by the multiple testing adjustment method, such as the procedure proposed by Benjamini and Hochberg (1995).

3 Application to Mouse Influenza Infection Data

In this section, we apply the proposed pipeline to the time course microarray data measured in mouse lungs after a nonlethal infection with influenza A PR8 (H1N1) (Pommerenke et al. 2012). The genome-wide changes in gene expression patterns were monitored at day 1, 2, 3, 5, 8, 10, 14, 18, 22, 26, 30, 40, and 60

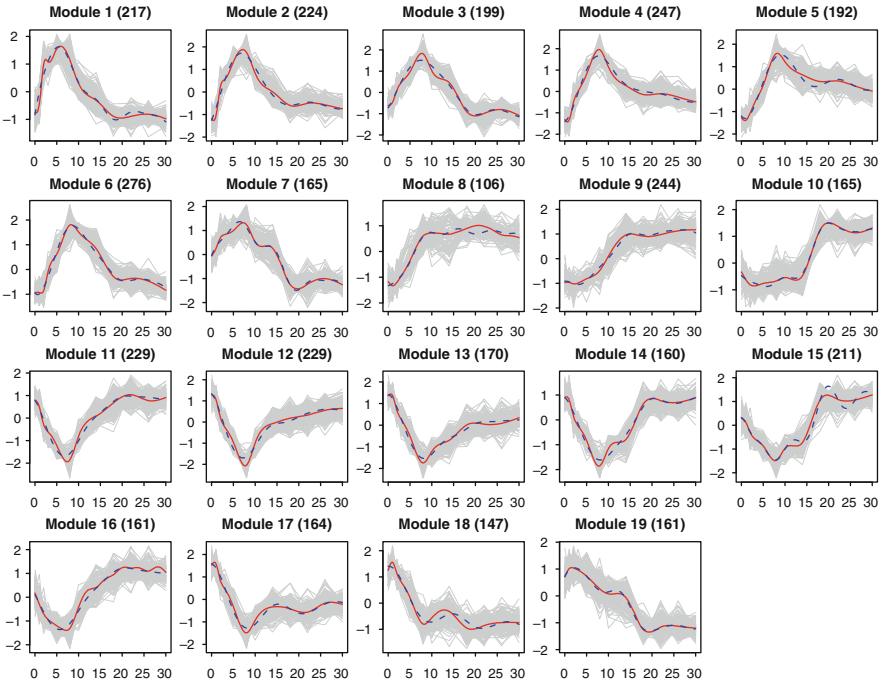


Fig. 1 The spaghetti plot of the standardized gene expression profiles for each functional module overlaid with the smoothed mean expression curve obtained from (6) (solid line) and the refined estimate from the linear ODE model (dashed line). The number of genes in each module is displayed in the parentheses.

post-infection (p.i.). Three mice were prepared as three independent biological replicates at each day p.i., except for days 3 and 5, where there were 6 replicates. Nine mice were mock-infected and their gene expression data were collected as baseline measurements (day 0). The total number of genes in the data set is 27527. In the following analysis, we only consider the gene expression data up to day 30 p.i. Since most of the gene expression profiles after day 30 p.i. are nearly flat, we believe that the majority of the gene expression variations are still preserved after removing the data on days 40 and 60 p.i.

We first perform the significance test under the null hypothesis (4). Using 10,000 permutations, we identify 3667 genes (FDR $= 0.005$) with temporally differential expression patterns. To further reduce the dimension of the problem, we group these differential genes into 19 functional modules using K-means clustering. Figure 1 shows the spaghetti plot of the standardized gene expression profiles for each functional module, with the number of genes in each module displayed in the parentheses. Treating the gene expression profiles within the same functional module as longitudinal data, we obtain the smoothed mean expression curve and the corresponding first order derivative for each module using model (6). The mean

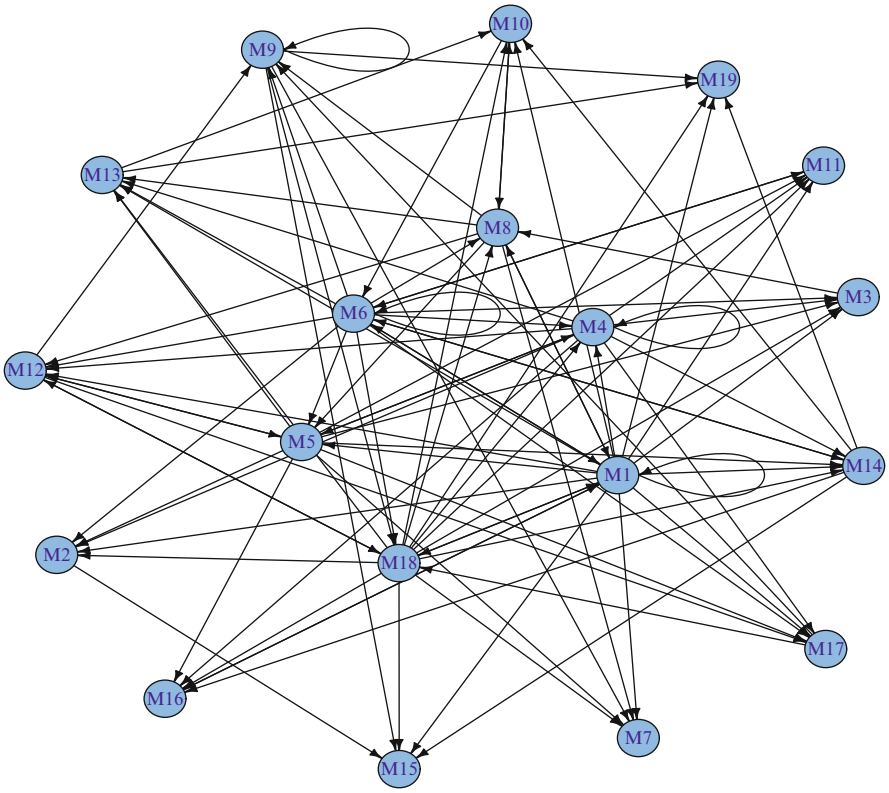


Fig. 2 The module-based gene regulatory network constructed by the linear ODE model from the viral infection gene expression data.

expression curves are displayed as solid lines overlaid with the spaghetti plots in Fig. 1, and as suggested by Liang and Wu (2008), these curve estimates are under-smoothed. We decouple the ODE system by plugging in the estimates of the mean expression curves and their derivatives, where the mean curve estimates and their derivatives are evaluated at an augmented time grid, with $N = 100$ time points equidistantly distributed in $[0, 30]$. Applying the SCAD variable selection method described in Sect. 2.3.2, we obtain the module-based GRN shown in Fig. 2. Conditional on the selected model structure, we refine the parameter estimates using the NLS method and the solutions from the refined ODE model are displayed as dashed lines in Fig. 1. We can see that these refined curve estimates from the linear ODE model closely follow the mean expression trend for each functional module.

Table 1 summarizes the detailed information about the inward and outward regulations between the functional modules. The negative signs in the table indicate down regulations, i.e., negative coefficients in the refined ODE model. We find that Modules 1, 4, 5, 6, 8, and 18 have the most outward regulations, indicating their crucial roles in this regulatory network. So we refer to them as the "hub"

Table 1 The inward and outward regulations, and the GO biological processes, KEGG pathways and selective molecular functions enriched for each functional module in the reconstructed regulatory network. The negative sign indicates a down regulation (a negative coefficient in the linear ODE model); otherwise it is an up regulation.

Module	Inward Influence	Outward Influence	Functional Annotation
M1	1, 6, 8^-, 16, 18^-	1, 2, 3,4, 5, 6, 7, 8, 10, 11^-, 12^-, 13^-, 14^-, 15^-, 16^-, 17, 18, 19^-	innate immune response, response to virus, inflammatory response, toll-like receptor signaling pathway, cytosolic DNA-sensing pathway,
M2	1, 4, 5, 6^-, 18	15	positive regulation of cytokine production, positive regulation of adaptive immune response, defense response cytokine-cytokine receptor interaction
M3	1, 4^-, 5, 6^-, 18	8^-	cellular metabolic process, nucleotide binding
M4	1, 4, 5^-, 6^-, 18^-	2, 3^-, 4, 5^-, 11, 12^-, 13, 14, 16, 17^-	natural killer cell mediated cytotoxicity, regulation of lymphocyte activation, regulation of leukocyte activation, defense response, T cell activation
M5	1, 4^-, 6^-, 8, 12^-	2, 3, 4^-, 7, 11, 12, 13, 14^-, 16^-, 17	T cell differentiation and activation, leukocyte and lymphocyte activation, hemopoietic or lymphoid organ development, hemopoiesis
M6	1, 6^-, 10, 11, 14^-	1, 2^-, 3^-, 4^-, 5^-, 6^-, 8, 9, 11, 12, 13, 14, 17, 18	M phase, cell cycle phase, cell division, mitotic cell cycle, mitosis
M7	1, 5, 8^-, 9, 18^-		ncRNA metabolic process, proteasome, tRNA metabolic process
M8	1, 3^-, 6, 10^-, 18^-	1^-, 5, 7^-, 9^-, 10, 12^-, 13^-	regulation of lymphocyte and leukocyte activation, B cell activation
M9	6, 8^-, 9, 12, 17^-	7, 9, 15, 18^-, 19^-	intestinal immune network for IgA production, asthma, B cell receptor signaling pathway
M10	1, 8, 13^-, 14, 18	6, 8^-	ribosome, regulation of transcription, DNA-dependent

M11	1^-, 4, 5, 6, 18	6	cellular process, development process
M12	1^-, 4^-, 5, 6, 8^-, 18^-	5^-, 9, 17, 18	microtubule-based process, motor activity, catalytic activity
M13	1^-, 4, 5, 6, 8^-, 18	10^-, 19	vascular smooth muscle contraction
M14	1^-, 4, 5^-, 6, 18^-	6^-, 10, 15, 16^-, 19^-	organ development, system development, cell differentiation, protein binding
M15	1^-, 2, 9, 14, 18		binding, system development, anatomical structure morphogenesis and development, organ development
M16	1^-, 4, 5^-, 14^-, 18^-	1	epithelial tube morphogenesis
M17	1, 4^-, 5, 6, 12	9^-, 18^-	drug metabolism, monooxygenase activity
M18	1, 6, 9^-, 12, 17^-	1^-, 2, 3, 4^-, 7^-, 8^-, 10, 11, 12^-, 13, 14^-, 15, 16^-, 19^-	catalytic activity
M19	1^-, 9^-, 13, 14^-, 18^-		ligase activity, negative regulation of protein metabolic process

modules. To better understand the roles these modules play in the regulatory network, we carry out functional enrichment analysis using DAVID (Huang et al. 2009). Due to space limitation, we only display the GO biological processes, the KEGG pathways and some selective molecular functions enriched in each functional module (Table 1). We can see that several modules have functions related to the immune process, such as innate immune response, positive regulation of adaptive immune response, T cell activation, and B cell receptor signaling pathway. The complete list of biological functions enriched in each module and also the genes associated with each functional term is available upon request.

Taking a closer look at Table 1, we find that genes in Module 1 are mostly related to the innate immune process. This is consistent with the expression patterns in Fig. 1, where genes respond to the viral infection immediately with rapid increases up to around day 6 p.i. and then quickly drop back to the baseline levels at around day 15 p.i. Modules 4 and 5 feature strong augmentations of gene expressions starting at around day 3 p.i., with peaks around day 8 p.i. (Fig. 1). Unlike Module 1, the gene expression levels after day 20 p.i. of Modules 4 and 5 are still considerably higher than those at the baseline. These two modules are mainly associated with the activation and differentiation of T cells and lymphocyte. B cell related functions are enriched in Modules 8 and 9, indicating the recruitment of B cells into the lungs. Accordingly, we find that the gene expressions of these two modules start to increase after day 5 p.i. and continue to be highly expressed till day 30 p.i. (Fig. 1). Modules 11–19 have down-regulated gene expression patterns, and interestingly, we find that the enriched functions in these modules are mainly housekeeping biological functions, such as cellular process, development, and binding. These down-regulated expressions reflect the death of the lung epithelial cells or their impairment due to the virus infection. The regulatory interactions between these housekeeping functions and the innate and adaptive immune responses enable the proper functioning of the immune process in lungs.

Besides the "intra-module" functional annotations, we are also interested in building a functional landscape of the genome-wide regulatory network in response to viral infection. Linking the functional annotations and the topology of the gene network identified through the ODE model (Fig. 2) can help us understand the functional linkages and associations between these functional modules, and thus better understand the dynamic regulations of the whole immune system. Taking Module 1 as an example, we can see from Table 1 that it is regulated by Modules 6, 8, 16, 18, and itself. Module 1 is associated with the innate immune response, while Modules 6, 8, 16, and 18 are enriched as "cell cycle," "B cell activation and immune response," "cell development," and "catalytic activity," respectively (Table 1). This means that many fundamental biological processes in cells are involved in the initiate of the innate immune response. On the outward influence side, we find Module 1 regulates almost all the other modules, indicating the crucial role of this module in the immune system.

In summary, the functional modules can be considered as meta-genes and the "intra-module' functional annotations summarize the biological functions of these meta-genes. The "inter-module" regulatory relationships imply the cooperative

coordination of these network units and the dynamics of the immune responses after the viral infection. We have shown in this data application that using the proposed pipeline, we are able to identify many interesting interactions in the mouse immune system and the regulatory relationships identified in the ODE-based GRN are consistent with those found by other investigators (Pommerenke et al. 2012).

4 Discussion

We proposed a new pipeline for the reverse-engineering of the dynamic GRN based on ODE models. We focused on the genome-wide time course gene expression data, which provides a complete view of the evolvement of the biological phenomena over a period of time rather than at a snapshot. Our pipeline is consisted of five steps: (1) detection of temporally differentially expressed genes; (2) clustering differential genes into functional modules; (3) identification of network structure; (4) parameter estimate refinement; and (5) functional enrichment analysis. A series of advanced statistical techniques are employed, including the integration of the FPCA and the hypothesis testing, time course gene expression clustering, nonparametric mixed-effects modeling and the parameter estimation and statistical inference for ODE models. The proposed pipeline is a computationally efficient data-driven tool bridging the experimental data and the mathematical modeling and statistical analysis. More importantly, the systems biological approaches in the pipeline allow us to model the living systems as a whole rather than a collection of single biological entities and gain insights into the control of a part of the system while taking into account the effect it has on the whole system. The application to the influenza-infected mouse data has illustrated the usefulness of the proposed pipeline.

For the identification of network structure, we adopted the two-stage smoothing-based estimation method to decouple the ODE system to a set of pseudo-regression models. Wu et al. (2012) proposed to use a weight function in the least squares criterion in the second stage after plugging in the estimates of the mean expression curves and their derivatives. They have shown that using the appropriate weight function can alleviate the boundary effects of the nonparametric curve estimates and also improve the convergence rate of the parameter estimates. Following this idea, we can modify the objective function (8) in the SCAD procedure to a penalized weighted least squares: $\frac{1}{2N}\sum_{j=1}^{N} d_k(t_j)(y_{kj} - z_i^T\beta)^2 + \sum_{i=1}^{K} J_{\lambda_k}(|\beta_{ki}|)$, where $d_k(t)$ is a nonnegative weight function on $[t_0, T]$ with boundary conditions $d_k(t_0) = d_k(T) = 0$.

An important extension of the proposed pipeline is to include some prior information in the ODE network modeling. For example, if one network component is known to regulate another component from the literature, one can incorporate this information in the network structure identification by not penalizing the corresponding regulatory coefficients in the variable selection procedure. Besides the popularly adopted linear ODE model (3), certain nonlinear ODE models are also very important in practical applications, because complex dynamic behaviors

and regulations in a GRN cannot be explained by simple linear systems (Heinrich and Schuster 1996). An example of nonlinear ODE models based on the functional modules can be written as

$$M_k'(t) = \beta_{k0} + \sum_{i=1}^{K} \beta_{ki} f_{ki}(M_i(t), \theta_{ki}), \quad k = 1, \ldots, K. \tag{10}$$

The regulatory functions f_{ki} can be known nonlinear functions with unknown parameters θ_{ki}, such as the sigmoid function model (Chen et al. 2004). They can also be some unknown functions and may be estimated using nonparametric techniques. Since co-regulations occur very often in GRNs, we can generalize model (10) to include the interaction effects between certain components. A further extension to (10) is to relax the additive assumption and allow more complex nonlinear relationships, such as the S-systems (Kikuchi et al. 2003; Kimura et al. 2005). However, these topics are beyond the scope of this paper and will be studied in the future research.

Acknowledgements This research was partially supported by the NIH grants HHSN 272201000055C, AI087135, and the University of Rochester CTSI pilot award (UL1RR024160) from the National Center For Research Resources.

References

BANSAL, M., GATTA, G. and DI BERNARDO, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics* **22** 815–822.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **57** 289–300.

CHEN, H.-C., LEE, H.-C., LIN, T.-Y., LI, W.-H. and CHEN, B.-S. (2004). Quantitative characterization of the trancriptional regulatory network in the yeast cell cycle. *Bioinformatics* **20** 1914–1927.

CHEN, T., HE, H. and CHURCH, G. (1999). Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing* 29–40.

DE JONG, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology* **9** 67–103.

DELYON, B., LAVIELLE, M. and MOULINES, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics* **27** 94–128.

EISEN, M., SPELLMAN, P., BROWN, P. and BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science USA* **95** 14863–14868.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

FRIEDMAN, N., LINIAL, M., NACHMAN, I. and PE'ER, D. (2000). Using bayesian networks to analyze expression data. *Journal of Computational Biology* **7 (3–4)** 601–620.

HARTIGAN, J. and WONG, M. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **28** 100–108.

HECKER, M., LAMBECKA, S., TOEPFERB, S., SOMERENC, E. and GUTHKE, R. (2009). Gene regulatory network inference: Data integration in dynamic models-a review. *BioSystems* **96** 86–103.

HECKERMAN, D. (1996). A tutorial on learning with bayesian networks. Tech. rep., Microsft Research.

HEINRICH, R. and SCHUSTER, S. (1996). *The regulation of cellular systems*. Chapman and Hall.

HIROSE, O., YOSHIDA, R., IMOTO, S., YAMAGUCHI, R., HIGUCHI, T., CHARNOCK-JONES, D. S., PRINT, C. and MIYANO, S. (2008). Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics* **24** 932–942.

HOLTER, N. S., MARITAN, A., CIEPLAK, M., FEDOROFF, N. V. and BANAVAR, J. R. (2001). Dynamic modeling of gene expression data. *Proceedings of the National Academy of Science USA* **98** 1693–1698.

HUANG, D., SHERMAN, B. and LEMPICKI, R. (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols* **4** 44–57.

JEONG, H., TOMBOR, B., ALBERT, R., OLTVAI, Z. and BARABÁSI, A.-L. (2000). The large-scale organization of metabolic networks. *Nature* **407** 651.

KIKUCHI, S., TOMINAGA, D., ARITA, M., TAKAHASHI, K. and TOMITA, M. (2003). Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics* **19** 643–650.

KIM, Y., CHOI, H. and OH, H.-S. (2008). Smoothily clipped absolute deviation on high dimensions. *Journal of the American Statistical Association* **103** 1665–1673.

KIMURA, S., IDE, K., KASHIHARA, A., KANO, M., HATAKEYAMA, M., MASUI, R., NAKAGAWA, N., YOKOYAMA, S., KURAMITSU, S. and KONAGAYA, A. (2005). Inference of s-system models of genetic networks using a cooperative coevolutionary algorithm. *Bioinformatics* **21** 1154–1163.

KOHONEN, T. (1997). *Self-Organizing Maps*. Springer, New York.

KOJIMA, K., YAMAGUCHI, R., IMOTO, S., YAMAUCHI, M., NAGASAKI, M., YOSHIDA, R., SHIMAMURA, T., UENO, K., HIGUCHI, T., GOTOH, N. and MIYANO, S. (2009). A state space representation of var models with sparse learning for dynamic gene networks. *Genome Informatics* **22** 56–68.

KUHN, E. and LAVIELLE, M. (2004). Coupling a stochastic approximation version of EM with a MCMC procedure. *ESAIM: Probability and Statistics* **8** 115–131.

LIANG, H. and WU, H. (2008). Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association* **103** 15701583.

LU, T., LIANG, H., LI, H. and WU, H. (2011). High dimensional ODEs coupled with mixedeffects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association* **106** 1242–1258.

LUAN, Y. and LI, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics* **19** 474–482.

MA, P., CASTILLO-DAVIS, C., ZHONG, W. and LIU, J. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research* **34** 1261–1269.

PERRIN, B., RALAIVOLA, L., MAZURIE, A., BOTTANI, S., MALLET, J. and D'ALCHÉ BUC, F. (2003). Gene networks inference using dynamic bayesian networks. *Bioinformatics* **19 (Suppl. 2)** ii138–148.

POMMERENKE, C., WILK, E., SRIVASTAVA, B., SCHULZE, A., NOVOSELOVA, N., GEFFERS, R. and SCHUGHART, K. (2012). Global transcriptome analysis in influenza-infected mouse lungs reveals the kinetics of innate and adaptive host immune responses. *PLoS One* **7** e41169.

SHMULEVICH, I., DOUGHERTY, E. R., KIM, S. and ZHANG, W. (2002). Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* **18** 261–274.

STEUER, R., KURTHS, J., DAUB, C. O., WEISE, J. and SELBIG, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* **18** S231–S240.

STUART, J. M., SEGAL, E., KOLLER, D. and KIM, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302** 249–255.

THOMAS, R. (1973). Boolean formalization of genetic control circuits. *Journal of Theoretical Biology* **42** 563–585.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58** 267–288.

VOIT, E. O. and ALMEIDA, J. (2004). Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics* **22** 1670–1681.

WU, H., XUE, H. and KUMAR, A. (2012). Numerical discretization-based estimation methods for ordinary differential equation models via penalized spline smoothing with applications in biomedical research. *Biometrics* **68** 344–352.

WU, H. and ZHANG, J.-T. (2005). *Nonparametric regression methods for longitudinal data analysis*. Wiley, New York.

WU, S. and WU, H. (2013). More powerful significant testing for time course gene expression data using functional principal component analysis approaches. *BMC Bioinformatics* **14** 6.

YEUNG, M. K. S., TEGNER, J. and COLLINS, J. J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Science USA* **99** 6163–6168.

ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics* **36** 1567–1594.

ZOU, M. and CONZEN, S. (2005). A new dynamic bayesian network approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* **21** 71–79.

Kernel Methods for Regression Analysis of Microbiome Compositional Data

Jun Chen and Hongzhe Li

Abstract With the development of next generation sequencing technologies, the human microbiome can now be studied using direct DNA sequencing. Many human diseases have been shown to be associated with the disorder of the human microbiome. Previous statistical methods for associating the microbiome composition with an outcome such as disease status focus on the association of the abundance of individual taxon or their abundance ratios with the outcome variable. However, the problem of multiple testing leads to loss of power to detect the association. When individual taxon-level association test fails, an overall test, which pools the individually weak association signal, can be applied to test the significance of the effect of the overall microbiome composition on an outcome variable. In this paper, we propose a kernel-based semi-parametric regression method for testing the significance of the effect of the microbiome composition on a continuous or binary outcome. Our method provides the flexibility to incorporate the phylogenetic information into the kernels as well as the ability to naturally adjust for the covariate effects. We evaluate our methods using simulations as well as a real data set on testing the significance of the human gut microbiome composition on body mass index (BMI) while adjusting for total fat intake. Our result suggests that the gut microbiome has a strong effect on BMI and this effect is independent of total fat intake.

J. Chen
Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA
e-mail: jchen@hsph.harvard.edu

H. Li (✉)
Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
e-mail: hongzhe@upenn.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_16,

1 Introduction

The human body carries ten times more bacterial cells than the human cells. Besides bacteria, there are other types of microbes such as viruses, fungi, and archaea. The collection of these microorganisms, their genetic elements, and their environmental interactions constitutes the human microbiome, which contains two orders of magnitude more genes than the genes in the human genome [1]. The human microbiome coevolves with the human genome and provides us with additional biological functions that are not encoded in the human genome. It is not until very recently that researchers began to appreciate the importance of human microbiome in human health and disease. With technological advancement, there have been studies of the human microbiome at different body sites ranging from skin [2] to gut [3–6]. Important insights have been gained from analysis of large-scale human microbiome data, including the discovery of enterotypes [4] and discovery of the link between diet and these enterotypes [6]. Meanwhile, many diseases have been linked to the disorder of the human microbiome including obesity, inflammatory bowel disease, and even colon cancers [7].

Prior to the next generation sequencing (NGS) era, the study of the human microbiome involves cloning individual microbial DNA fragments followed by Sanger sequencing, a procedure that is very laborious and expensive. NGS technologies such as 454 pyrosequencing and Illumina Solexa sequencing have enabled researchers to study the human microbiome using direct DNA sequencing, which is much cheaper, faster, and more high-throughput. There are basically two ways of sequencing the microbiome: the 16S rRNA gene targeted sequencing and the shotgun metagenomic sequencing [8]. The latter involves sequencing all the microbial DNA fragments from a microbiome sample. Though the shotgun method is more powerful and possible less biased, the 16S rRNA gene targeted approach is routinely conducted to determine the bacterial composition due to its relative low cost and simpler analysis pipeline. In the targeted approach, one variable region of the 16S rRNA gene is isolated by PCR and sequenced by NGS. Since each bacterial species harbors different version of 16S sequence, these sequenced 16S tags carry information about species identity. The 16S tags from the same species are highly similar, so in practice the 16S tags are clustered into small units called Operational Taxonomic Units or OTUs at 97 % similarity level. These OTUs are considered to be surrogates of biological species, though strict correspondence is hard to establish. The number of OTUs is usually on the order of thousands or more and varies tremendously depending on the body sites studied and the number of samples sequenced.

The OTUs are related by a phylogenetic tree. This tree provides an important prior information on the phylogenetic relationships among the bacterial species. Appropriate use of the tree information can lead more meaningful analysis [9, 10, 12]. Figure 1 gives an example of the OTU counts generated for a study on the effects of diet on the gut microbiome composition [6]. The phylogenetic tree, shown on the top, relates all the OTUs and exhibits clear clustering pattern.

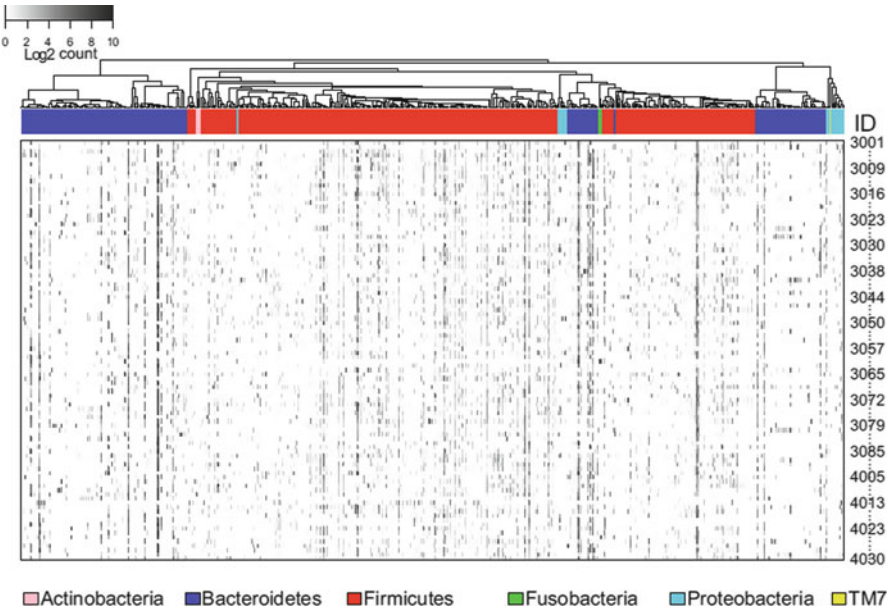


Fig. 1 OTU counts of 98 gut microbiome samples from COMBO study. Row: subjects, column: OTUs. The gray scale indicates the OTU abundance. White regions indicate the OTUs that receive zero counts in a given sample. The OTUs are related by a phylogenetic tree shown at the top. The tree is constructed using the genetic distances between OTUs based on the aligned 16S sequences. The OTUs from different phyla are colored accordingly.

Understanding the relationship between the human microbiome and disease is one of the most important goals of human microbiome studies. Testing the significance of the effects of overall microbiome composition on a disease status or a disease related quantitative trait is an essential step to disentangle the complex relationship between the human microbiome and disease. However, due to the complexity of the microbiome data, rigorous statistical methods are required to conduct such an overall test. An overall test can overcome the problem of multiple testing of individual associations between microbial taxon and disease that are usually very weak. In this paper, we propose a kernel-based semi-parametric regression method for testing the significance of the effect of overall microbiome composition on an outcome variable. The key of our approach is to define a distance matrix among the samples that effectively take into account the observed OTU compositions and the phylogenetic tree information. The phylogenetic tree-based distance matrix can then be used to define a kernel function used in semi-parametric generalized linear models. A score test can then be developed to test for an overall association between the microbiome composition and an outcome variable.

We briefly review the literature on kernel-based regression and its connection with generalized linear mixed-effects models. We then present methods for kernel construction for microbiome compositional data using phylogenetic tree information. We finally present simulation results and real data application.

2 Methods

2.1 Kernel-Based Semi-parametric Regression Model and Score Test

Suppose we have n samples. Let y_i denote an outcome variable for the ith sample, where y_i can be binary (e.g., disease status) or continuous (e.g., a quantitative trait). Let $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})$ be the p covariates we want to adjust for and $\mathbf{z}_i = (z_{i1}, \cdots, z_{iq})$ be the abundances of q OTUs for the ith sample, where q is large and is usually in the order of thousands. These OTUs are related by a known phylogenetic tree $\mathscr{T}$. To test the effect of OTUs on the outcome after adjusting for the covariate effects, we use the kernel-based semi-parametric regression framework [13, 14], in which we model the effects of OTUs and other covariates using the following link function $g(.)$,

$$g(\mu_i) = \mathbf{x}_i^T \beta + h(\mathbf{z}_i),$$

where μ_i is the mean of the outcome, the covariate effects are modeled parametrically with β being the regression coefficients, and the OTU effects are modeled nonparametrically. Specifically, we assume that the general function $h(\mathbf{z}_i)$ is from a Reproducing Kernel Hilbert Space $\mathscr{H}_K$ generated by a positive semi-definite kernel function $K(.,.)$. Under this framework, testing no OTU effect on the outcome is equivalent to testing $h(\mathbf{z}) = 0$. When the outcome is a continuous normal variable, $g(.)$ is the identity function. When the outcome is binary variable, $g(.)$ is the logit function $\ln(\mu_i/(1-\mu_i))$. By choosing different kernel functions, we assume different functional form of $h(.)$, so we can model OTU interactions, nonlinear OTU effects, and even the phylogenetic relationship among OTUs.

The regression coefficients β and the function $h(.)$ can be estimated by maximizing the penalized likelihood

$$pl(h,\beta) = \sum_{i=1}^{n} l(h,\beta; y_i, \mathbf{x}_i, \mathbf{z}_i) - \frac{1}{2}\lambda \|h\|^2_{\mathscr{H}_K},$$

where $l(.)$ is the likelihood function, $\|h\|_{\mathscr{H}_K}$ measures the smoothness of the function $h(.)$, and the tuning parameter λ controls the trade-off between model fit and complexity of the function $h(.)$. By representer theorem, the penalized maximum likelihood estimate of $h(.)$ can be expressed as $h(.) = \sum_{i=1}^{n} \alpha_i K(., \mathbf{z}_i)$ for some α. So the penalized likelihood function can be rewritten as

$$pl(\alpha,\beta) = \sum_{i=1}^{n} l(\alpha,\beta; y_i, \mathbf{x}_i, \mathbf{z}_i) - \frac{1}{2}\lambda \alpha^T \mathbf{K} \alpha,$$

where $\mathbf{K}$ is the kernel matrix over the samples with $K_{ij} = K(\mathbf{z}_i, \mathbf{z}_j)$. This penalized likelihood function is equivalent to the Penalized Quasi-likelihood (PQL) function from a generalized mixed-effects model:

$$g(\mu_i) = \mathbf{x}_i^T \beta + h_i,$$

where β are fixed effects and $\mathbf{h} \sim N(\mathbf{0}, \tau\mathbf{K})$ are subject-specific random effects with the variance component $\tau = 1/\lambda$ and $\mathbf{h} = \mathbf{K}\alpha$. Through this connection, parameter estimation and testing in original kernel-based regression framework can be conducted under the generalized mixed-effects model. Testing $h(\mathbf{z}) = 0$ corresponds to the variance component test in generalized mixed model:

$$H_0 : h(\mathbf{z}) = 0 \Leftrightarrow H_0 : \tau = 0.$$

The variance component test can be performed using a score test and the score statistic is

$$Q = \frac{1}{2\phi}(\mathbf{y} - \hat{\mu}_0)^T \mathbf{K}(\mathbf{y} - \hat{\mu}_0), \tag{1}$$

where ϕ is the dispersion parameter, $\hat{\mu}_0$ is the fitted mean under H_0. The distribution of Q under H_0 has a weighted mixture of χ^2 distributions and can be determined using approximation or exact methods. Details of the score test can be found in [14]. The power of the score test depends on the choice of the kernel matrix $\mathbf{K}$. In the next section, we provide a way of designing an appropriate kernel that takes into account the phylogenetic relationship among the OTUs.

2.2 Construct Kernels Based on Generalized UniFrac Distances

The kernel-based score test requires a kernel matrix K. One major characteristic of the OTU data is that the OTUs are related by a phylogenetic tree. OTUs are usually clustered on the tree and form lineages of different depths. These lineages can be roughly thought as taxonomic units at different levels such as genus and family. OTUs from the same lineage are genetically more similar and usually have similar biological functions. The outcome is usually affected by one or several OTU lineages instead of dispersed individual OTUs. One simple way of defining a kernel matrix is by transforming the distance matrix through

$$\mathbf{K} = -\frac{1}{2}(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n})\mathbf{D}^2(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}), \tag{2}$$

where $\mathbf{D} = \{d_{ij}\}$ is the pairwise distance matrix, $\mathbf{I}$ is the identity matrix, and $\mathbf{1}$ is a vector of 1's. It is easy to verify that kernel matrix defined this way can recover the original distances using standard kernel operation: $d_{ij}^2 = K_{ii} + K_{jj} - 2K_{ij}$. If the $\mathbf{D}$ is Euclidean, *i.e.* the samples can be embedded in a real coordinate space, then the positive semi-definiteness of $\mathbf{K}$ is guaranteed. If there is no Euclidean embedding for $\mathbf{D}$, then $\mathbf{K}$ is not positive semi-definite. In such case, we can apply a positive semi-definiteness correction procedure. We first perform an eigen-decomposition by $\mathbf{K} = \mathbf{U}\Lambda\mathbf{U}^T$, where $\Lambda = Diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$. We then reconstruct $\mathbf{K}^*$ using the absolute eigenvalues by $\mathbf{K}^* = \mathbf{U}\Lambda^*\mathbf{U}^T$, where $\Lambda^* = Diag(|\lambda_1|, |\lambda 2|, \ldots, |\lambda_n|)$.

The most widely used distance metric in microbiome data analysis is UniFrac distance that effectively incorporates the phylogenetic tree information [15]. The unweighted UniFrac distance, which uses only OTU absence and presence information, is defined as the fraction of branch length of the phylogenetic tree that is unique to any microbial community. The weighted UniFrac distance uses OTU proportion data and weights the branch length by proportion difference of the two communities. A generalized UniFrac distance was also proposed [10] that can be regarded as a one-parameter extension of the UnFrac distances. The generalized UniFrac distance between communities A and B is defined as:

$$d^{(\xi)} = \frac{\sum_{i=1}^{m} b_i (p_i^A + p_i^B)^{\xi} \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{m} b_i (p_i^A + p_i^B)^{\xi}},$$

where m is the number of branches in the tree, b_i is the length of ith branch, and p_i^A, p_i^B are the proportions of OTU lineages that descend from ith branch for community A and B. The term on the denominator is the normalizing factor so $d^{(\xi)} \in [0,1]$. In the formula, the branch length is weighted by both the relative difference $\left|p_i^A - p_i^B\right| / \left|p_i^A + p_i^B\right|$, which puts equal emphasis on every branch and the branch proportion $(p_i^A + p_i^B)^{\xi}$, which includes a parameter ξ controlling the weight on abundant lineages. As ξ changes from 0 to 1, more weight is put on abundant lineages. The generalized UniFrac distance includes weighted and unweighted UniFrac distances as special cases. When $\xi = 1$, the generalized UniFrac distance is reduced to weighted UniFrac distance d^W. When $\xi = 0$ and we use the OTU absence and presence data, the generalized UniFrac distance is reduced to unweighted UniFrac distance d^U. By taking different values of ξ, we can detect a much wider range of community difference, from the difference in most abundant lineages to least abundant lineages. To avoid the multiple testing problem using too many ξ's, we restrict us to four representative instances on the generalized UniFrac distance series, namely d^W, $d^{(0.5)}$, $d^{(0)}$, and d^U. Note that $d^{(0)}$ uses the proportion data and is different from d^U.

We use the kernels constructed from these four distances to detect microbiome effects using both simulations and a real application. Specifically, for two individuals i and j, we calculate their pair-wise distance, d_{ij}^W, $d_{ij}^{(0.5)}$, $d_{ij}^{(0)}$, and d_{ij}^U, respectively, which defines the distance matrix and the kernel function as in Eq. (2).

3 Simulations

We present simulations to evaluate the power of kernel-based score tests for detecting the microbiome effects on a disease status or a quantitative trait. We compare the performance of different kernels under various scenarios to identify

the best-performance scenario for each kernel considered. For each scenario, we generate 100 samples and the power is calculated based on 1,000 repetitions at type I error 0.05.

3.1 Simulation Methods

To mimic the real OTU counts, we use a medium-scale phylogenetic tree of OTUs from a real throat microbiome data set of 60 samples [16]. The phylogenetic tree, which consists of 856 OTUs after discarding the singleton OTUs, is constructed using the FastTree algorithm in the Qiime pipeline [17]. We model the OTU counts using Dirichlet-multinomial (DM) model to account for the overdispersion seen in real microbiome data [11]. We estimate the mean OTU proportions $(\pi_1, \pi_2, \cdots, \pi_{856})$ as well as the overdispersion parameter θ using maximum likelihood method. We then generate the OTU counts using the estimated parameters. Specifically, for ith sample, the OTU proportions are generated using Dirichlet distribution: $(p_{i1}, p_{i2}, \cdots, p_{i856}) \sim Dirichlet(\widehat{\pi}_1, \widehat{\pi}_2, \cdots, \hat{\pi}_{856}, \hat{\theta})$ and the OTU Counts are generated using a Multinomial Distribution with the OTU proportions generated in previous step: $(z_{i1}, z_{i2}, \cdots, z_{i856}) \sim Mutinomial(p_{i1}, p_{i2}, \cdots, p_{i856}, n_i)$, where the sequencing depth n_i is drawn from a negative binomial distribution with a mean 1,000 and size 25.

We partition the 856 OTUs into 20 clusters (lineages) by performing PAM (Partition Around Medoid) based on the OTU distance matrix, where the pair-wise OTU distance is taken to be the patristic distance, *i.e.* the length of the shortest path linking the two OTUs on the phylogenetic tree. The abundance of these 20 OTU clusters varies tremendously. We assume that the outcome variable depends on the abundance of a certain OTU cluster corresponding to some bacterial lineage. To compare the power of different kernels under different scenarios, we consider in our simulation four representative clusters with different abundance levels (19.4%, 5.7%, 1.5%, 0.3%, respectively). The continuous outcome is generated using a normal distribution with mean $f \sum_{k \in S_j} p_{ik}$ and fixed standard deviation, where S_j is the set of OTU indices in cluster j and the parameter f controls the effects of the OTU cluster. The standard deviation is the same for different clusters (SD=1.0). Specifically, we first standardize the proportions of the OTU in the associated cluster to have mean 0 and standard deviation of 1. The binary outcome is generated using a Bernoulli distribution with the mean $1/(1+\exp(-f \sum_{k \in S_j} p_{ik}))$. We vary the effect size f to create power curves. When $f = 0$, the OTU cluster has no effects on the outcome. The four generalized UniFrac distances d^W, $d^{(0.5)}$, $d^{(0)}$, and d^U are constructed using GUniFrac package [9] with the observed OTU counts and the phylogenetic tree. We then convert the UniFrac distances into kernels using formula (2). The kernel matrix is then used in the score statistic (1) for testing the overall association. The power is calculated based on 1,000 repetitions with type I error rate 0.05.

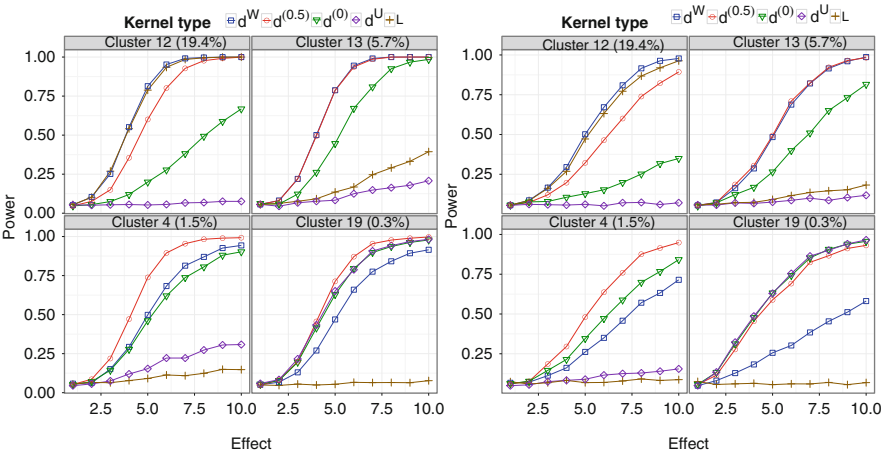


Fig. 2 Statistical power for the kernel-based score test with different kernels for testing the effects of the overall microbiome composition on an outcome. Four representative OTU clusters, ranging from very abundant to very rare, are chosen to affect the outcome. The abundance of corresponding cluster is shown in the parenthesis. The kernel-based test is used for testing the overall association of between all OTUs and the outcome. The power curve is created by varying the OTU effects. The first point of the power curve shows the power under null. Kernels compared are those constructed from generalized UniFrac distances (d^W, $d^{(0.5)}$, $d^{(0)}$ and d^U) and the linear kernel (L). Both continuous outcome (A) and binary outcome (B) are investigated.

3.2 *Simulation Results*

Figure 2 shows the power of kernel-based score test using kernels constructed from the generalized UniFrac distances d^W, $d^{(0.5)}$, $d^{(0)}$, and d^U and the simplest linear kernel, *i.e.* inner product of the OTU proportion vectors. Note that although the true associated OTUs belong to a cluster, the kernel-based tests are used for testing the overall association between all OTUs and the outcome. These different kernels behave qualitatively the same for both the continuous outcome (panel A) and binary outcome (panel B). Clearly, no kernel performs the best under all scenarios considered. When there is no OTU effect, score test controls the type I error at 0.05 level. Each kernel has its best performing scenario. The kernel based on d^W is most powerful to detect the effects of abundant OTU cluster (cluster 12) on the outcome, but it becomes less powerful as the causal OTU cluster becomes less abundant. The kernel based on d^U shows the opposite direction with its best performance for rare OTU cluster (cluster 19). The kernel based on $d^{(0)}$ lies in the middle. Most remarkably, the kernel based on $d^{(0.5)}$ is most robust with its performance always being closer to the best performing kernel for the corresponding scenario. It also achieves best power when the OTU cluster is moderately abundant (cluster 4) and is comparable to d^U when the OTU cluster is rare (cluster 19). In comparison, the linear kernel, which does not use phylogenetic tree information, is only powerful to detect the effects of abundant OTU cluster, where the kernel based on d^W achieves

a similar or better power, but it becomes almost powerless in other scenarios. In other words, the linear kernel is dominated by the kernel based on d^W. Therefore, incorporation of the phylogenetic tree information can increase the power of the score test. In practice, it is helpful to conduct score tests using different UniFrac distance-based kernels. The test results can provide potential insights into whether the rare or abundant OTUs are associated with the outcome.

4 Application to Real Data Analysis

To demonstrate the use of our proposed method in real applications, we apply our method to a real data set for testing the significance of the effect of the overall microbiome composition on the Body Mass Index (BMI), an index of obesity. Recent mouse and human studies have implicated the role of the gut microbiome in obesity [18–20]. However, these studies all focused on testing the association of obesity with particular microbiome features such as the species richness and the ratio of Firmicutes to Bacteroidetes. We use the microbiome data set from a study on the effects of diet on gut microbiome composition conducted at the University of Pennsylvania [6]. For this study, both gut microbiome 16S data and BMI data are available for 98 healthy subjects. Fecal samples were obtained from these 98 subjects and bacterial DNA was extracted using standard protocol. After multiplexed 454 pyrosequencing, about 900,000 high quality, partial (∼370bp) 16S rRNA gene sequences were generated. These sequences were analyzed using the Qiime pipeline [17], where the sequences were clustered at 97% sequence identity into OTUs. Generalized UniFrac distances (d^W, $d^{(0.5)}$, $d^{(0)}$ and d^U) are computed using the OTU abundance data and the OTU tree and are converted to kernels. We first perform kernel-based score test on the original continuous BMI values. Table 1 shows that the score test is most significant when the kernel based on d^U is used ($p = 6.7E-6$) and becomes less significant when more weights are put on abundant lineages (*e.g.* d^W), indicating it is mostly the rare lineages that affect the BMI. This is in accordance with the observation that the species richness decreased in obese twins [18], since these rare species contribute to the most of the species richness. To illustrate the kernel-based test for binary outcome, we dichotomized subjects into normal or overweight with a BMI cutoff value 25. The test results have the same trend.

Fat intake is a confounding factor in the test, since it is correlated with both BMI ($r = 0.32, p = 0.001$) and microbiome composition (PERMANOVA, $p < 0.05$).

Table 1 Kernel-based score test for the effect of the overall gut microbiome composition on BMI (unadjusted).

Kernel type	d^W	$d^{(0.5)}$	$d^{(0)}$	d^U
p-value (original scale)	0.22	0.03	4.3E-5	6.7E-6
p-value (dichotomized)	0.21	0.02	7.9E-5	4.2E-5

Table 2 Kernel-based score test for the effect of the overall gut microbiome composition on BMI after adjusting for fat intake and the effect of fat intake on BMI after adjusting for the microbiome composition.

Kernel type	d^W	$d^{(0.5)}$	$d^{(0)}$	d^U
p-value (microbiome)	0.36	0.11	0.007	0.003
p-value (fat)	0.001	0.006	0.012	0.036

We next conduct kernel-based score test after adjusting for fat intake and the result is shown in Table 2. Though the significance of the microbiome effect is reduced, it is still significant when the kernels based on $d^{(0)}$ and d^U are used ($p = 0.007, 0.003$ respectively), indicating the association of overall microbiome composition and BMI after adjusting for the total fat intake. Table 2 also gives the p-values for the fat intake adjusting for microbiome effect. It remains significant no matter what kernels are used.

Acknowledgments We thank Rick Bushman, James Lewis, and Gary Wu for sharing the data and for many helpful discussions. This research is supported by NIH grants CA127334 and GM097505.

References

[1] Cho I, Blaser MJ (2012) The human microbiome: at the interface of health and disease. Nat Rev Genet 13(4):260–270
[2] Grice EA, Kong HH, Conlan S et al (2009) Topographical and temporal diversity of the human skin microbiome. Science 324(5931):1190–1192
[3] Qin J, Li R, Raes J et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464(7285):59–65
[4] Arumugam M, Raes J, Pelletier E et al (2011) Enterotypes of the human gut microbiome. Nature 473(7346):174–180
[5] Muegge BD, Kuczynski J, Knights D et al (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science 332(6032):970–974
[6] Wu GD, Chen J, Hoffmann C et al (2011) Linking long-term dietary patterns with gut microbial enterotypes. Science 334(6052):105–108
[7] Kinross JM, Darzi AW, Nicholson JK (2011) Gut microbiome-host interactions in health and disease. Genome Med 3(3):14
[8] Kuczynski J, Lauber CL, Walters WA et al(2011) Experimental and analytical tools for studying the human microbiome. Nat Rev Genet 13(1):47–58
[9] Chen J, Bittinger K, Charlson ES et al (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 28(16):2106–2113
[10] Chen J, Bushman FD, Lewis JD et al (2012) Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. Biostatistics, doi: 10.1093/biostatistics/kxs038
[11] Chen J, Li H (2012) Variable Selection for Sparse Dirichlet-Multinomial Regression with An Application to Microbiome Data Analysis. Ann Appl Stat, in press
[12] Purdom E (2011) Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree. Ann Appl Stat 5(4):2326–2358

[13] Liu D, Lin X, Ghosh D (2007) Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. Biometrics 63(4):1079–1088
[14] Liu D, Ghosh D, Lin X (2008) Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC bioinformatics 9(1):292
[15] Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71(12):8228–8235
[16] Charlson ES, Chen J, Custers-Allen R et al(2010) Disordered microbial communities in the upper respiratory tract of cigarette smokers. PloS One 5(12): e15216
[17] Caporaso JG, Kuczynski J, Stombaugh J et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat methods 7(5):335–336
[18] Turnbaugh PJ, Hamady M, Yatsunenko T et al(2008) A core gut microbiome in obese and lean twins. Nature 457(7228):480484
[19] Hildebrandt MA, Hoffmann C, Sherrill-Mix SA et al (2009) High-fat diet determines the composition of the murine gut microbiome independently of obesity. Gastroenterology 137(5):1716–1724.
[20] Ley RE (2010) Obesity and the human microbiome. Curr opin gastroen 26(1):5.

A Conditional Autoregressive Model for Detecting Natural Selection in Protein-Coding DNA Sequences

Yu Fan, Rui Wu, Ming-Hui Chen, Lynn Kuo, and Paul O. Lewis

Abstract Phylogenetics, the study of evolutionary relationships among groups of organisms, has played an important role in modern biological research, such as genomic comparison, detecting orthology and paralogy, estimating divergence times, reconstructing ancient proteins, identifying mutations likely to be associated with disease, determining the identity of new pathogens, and finding the residues that are important to natural selection. Given an alignment of protein-coding DNA sequences, most methods for detecting natural selection rely on estimating the codon-specific nonsynonymous/synonymous rate ratios (d_N/d_S). Here, we describe an approach to modeling variation in the d_N/d_S by using a conditional autoregressive (CAR) model. The CAR model relaxes the assumption in most contemporary phylogenetic models, i.e., sites in molecular sequences evolve independently. By incorporating the information stored in the Protein Data Bank (PDB) file, the CAR model estimates the d_N/d_S based on the protein three-dimensional structure. We implement the model in a fully Bayesian approach with all parameters of the model considered as random variables and make use of the NVIDIA's parallel computing architecture (CUDA) to accelerate the calculation. Our result of analyzing an empirical abalone sperm lysine data is in accordance with the previous findings.

Y. Fan (✉)
Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, 1400 Pressler Dr., FCT4.6000, Houston, TX 77030, USA
e-mail: yfan1@mdanderson.org

R. Wu • M.-H. Chen • Lynn Kuo
Department of Statistics, The University of Connecticut, 215 Glenbrook Rd., Unit 4120, Storrs, CT 06269, USA
e-mail: rui.wu@uconn.edu; ming-hui.chen@uconn.edu; lynn.kuo@uconn.edu

P.O. Lewis
Department of Ecology and Evolutionary Biology, The University of Connecticut, 75 N. Eagleville Rd., Unit 3043, Storrs, CT 06269, USA
e-mail: paul.lewis@uconn.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_17,

1 Introduction

Darwinian natural selection at the molecular level is always an attractive topic among biologists because it is the basis of explaining adaptation in morphology, physiology, and developmental biology, and understanding species divergence and evolutionary innovations. By comparing protein-coding genes and differentiating synonymous substitutions from nonsynonymous substitutions, the traces of natural selection can be detected. The nonsynonymous/synonymous rate ratio ($\omega = d_N/d_S$) measures selective pressure: (1) $\omega = 1$, neutral substitution; (2) $\omega > 1$, positive selection; and (3) $\omega < 1$, purifying selection. Most contemporary phylogenetic substitution models assume that sites in molecular sequences evolve independently and identically (i.i.d.) for the convenience of likelihood calculation. But this assumption is not biologically plausible, for instance, substitutions of both RNA and protein-coding genes must be affected by their functional constraints that vary from site to site. Various substitution models were proposed to relax the i.i.d. assumption, for example, the auto-discrete-gamma model [1], the hidden Markov-chain model [2], the mechanistic model [3], and the phylogeny and secondary structure using maximum likelihood (PASSML) model [4, 5]. Besides those, Robinson et al. [6], Rodrigue et al. [7], and Kleinman et al. [8] put forward a Bayesian model that takes the whole sequence as the Markov chain state instead of every amino acid site. Huelsenbeck et al. [9] introduced a nonparametric Dirichlet process (DP) model that allowed the sites to be automatically assigned into one of a number of classes with each class having a different ω without changing the likelihood calculation.

We introduce here a conditional autoregressive (CAR) [10] model for the purpose of both relaxing the i.i.d. assumption in most substitution models and detecting positive selection in protein-coding genes. Moreover, the fast development of structural biology makes it possible to apply the CAR model to detecting positive selection since it provides the relevant spatial information.

2 Materials and Methods

2.1 Phylogenetic Codon Substitution Model with CAR Prior

On the assumption that substitutions follow a continuous-time Markov process on the phylogenetic tree, the instantaneous rate matrix **Q** of the general codon model can be expressed in Eq. (1), whose entries represent the instantaneous rates of changing from codon i to codon j,

$$\mathbf{Q} = \{q_{ij}\} = \begin{cases} a\omega\pi_j : \text{nonsynonymous transversion (A} <-> \text{C)} \\ a\pi_j : \text{synonymous transversion (A} <-> \text{C)} \\ b\omega\pi_j : \text{nonsynonymous transition (A} <-> \text{G)} \\ b\pi_j : \text{synonymous transition (A} <-> \text{G)} \\ c\omega\pi_j : \text{nonsynonymous transversion (A} <-> \text{T)} \\ c\pi_j : \text{synonymous transversion (A} <-> \text{T)} \\ d\omega\pi_j : \text{nonsynonymous transversion (C} <-> \text{G)} \\ d\pi_j : \text{synonymous transversion (C} <-> \text{G)} \\ e\omega\pi_j : \text{nonsynonymous transition (C} <-> \text{T)} \\ e\pi_j : \text{synonymous transition (C} <-> \text{T)} \\ f\omega\pi_j : \text{nonsynonymous transversion (G} <-> \text{T)} \\ f\pi_j : \text{synonymous transversion (G} <-> \text{T)} \\ 0 : i \text{ and } j \text{ differ at more than one position,} \end{cases} \quad (1)$$

where $\omega = d_N/d_S$ is the nonsynonymous/synonymous rate ratio, a, b, c, d, e, and f are the generalized time-reversible (GTR) exchangeabilities, and π_j is the stationary frequency of codon j. We let ω have a CAR prior. The dimension of the instantaneous rate matrix $\mathbf{Q}$ is 61×61 because the three stop codons are excluded from the state space. Every codon has its own instantaneous rate matrix $\mathbf{Q}$, and every $\mathbf{Q}$ is rescaled before the calculation of the transition probabilities, $\mathbf{P}(t) = e^{\mathbf{Q}t}$, which means that the branch lengths, t, of the phylogenetic tree are in terms of expected number of substitutions per codon.

The CAR prior is in the form of a multivariate normal distribution whose mean vector is μ and covariance matrix is Σ with K-by-K dimensions. K is equal to the number of ω in the data. In practice, the precision matrix Σ^{-1} is used instead of the covariance matrix Σ. The precision matrix is defined as

$$\Sigma^{-1} = \mathbf{D}_w - \rho\mathbf{W}, \quad (2)$$

where $\mathbf{W}$ is the proximity matrix whose entry w_{ij} equals to 1 or 0 based on the a priori knowledge that the codons i and j are neighbors or not, $\mathbf{D}_w$ is a diagonal matrix whose diagonal element is the summation of the corresponding row of $\mathbf{W}$, and ρ is a scalar whose domain belongs to $\left(1/\lambda_{(1)}, 1/\lambda_{(n)}\right)$, where $\lambda_{(1)} < \lambda_{(2)} < \cdots < \lambda_{(n)}$ are the ordered eigenvalues of $\mathbf{D}_w^{-1/2}\mathbf{W}\mathbf{D}_w^{-1/2}$ [11]. Because the nonsynonymous/synonymous rate ratio (ω) must be positive, we let $\omega = e^{\phi}$ and ϕ has the CAR prior. Therefore, the prior of ω is indeed a multivariate log-normal distribution, and the full probability density function (PDF) of ω is,

$$f(\omega|\rho,\tau) = \frac{\left|\Sigma^{-1}\right|^{\frac{1}{2}}\tau^{-K}}{(2\pi)^{\frac{K}{2}}\prod_{i=1}^{K}\omega_i} \cdot \exp\left\{-\frac{1}{2\tau^2}(\log\omega)'\Sigma^{-1}(\log\omega)\right\} \cdot I_{\{1/\lambda_{(1)}<\rho<1/\lambda_{(n)}\}}(\rho), \quad (3)$$

where τ is a scalar tuning the variance.

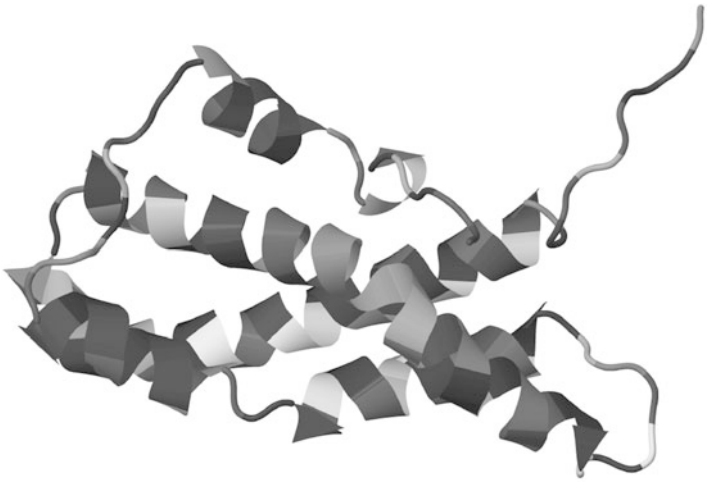


Fig. 1 Image showing the abalone sperm lysin based on the 2LIS PDB data file using Jmol-12.0.49. The cartoon-style is selected, and the image is colored by amino acids.

2.2 Empirical Data Analysis

The Protein Data Bank (PDB) that is supervised by the Worldwide Protein Data Bank (wwPDB) [14] is a repository for the 3-D structures of large biological molecules, such as proteins and nucleic acids. PDB data files can be retrieved from the web sites of RCSB PDB (USA), PDBe (Europe), and PDBj (Japan) that are the founding members of the wwPDB using file IDs or key words. A typical PBD data file contains a series of title sections, including the names of molecules, details about data collection and structure solution, primary and secondary structure information, atomic coordinates, crystallographic structure factors, NMR experimental data, bibliographic citations, etc. All the information is arranged by following the PDB format, which makes it possible to extract the atomic coordinates.

The abalone sperm lysin data set consisting of 25 taxa and 405 nucleotide sites [12, 13] was examined for the purpose of detecting positive selection in protein-coding DNA sequences. The corresponding PDB data file, ID:2LIS [15] (Fig. 1), was downloaded to specify the proximity matrix (**W**) of the CAR prior. Before the proximity matrix was determined, a matrix **S** whose entries represented the shortest distance between any two residues was calculated first in the light of the PDB data file,

$$\mathbf{S} = \{s_{ij}\} = \begin{cases} 0 & \text{if} \quad i = j \\ d_{ij,min} & \text{if} \quad i \neq j \end{cases}, \tag{4}$$

where i and j denoted the residue site, all the entries were in angstroms, and $d_{ij,min}$ was the minimum Euclidean distance of all possible calculations given all the

atomic orthogonal coordinates of any two residues were known. In total, four spatial correlation schemes were examined: (1) "Simple" (the simplest spatial correlation, i.e., the two adjacent residues would be treated as neighbors); (2) "Radius 5" (any two residues whose shortest distance (s_{ij}) was less or equal than 5 would be thought as neighbors); (3) "Radius 10" (the same as scheme (2) except that the bound was set to 10); and (4) "Radius 20" (the same as scheme (2) except that the bound was set to 20).

The general codon model (Eq. (1)) was applied to all analyses, and the tree topology was fixed to the best tree topology that were accompanied by the original data. Prior distributions were as follows: Dirichlet($\underbrace{1,\cdots,1}_{61}$) for the codon frequencies; Dirichlet(1,1,1,1,1,1) for the GTR exchangeabilities; Inverse-gamma(2.001,1) for the hyperparameter τ^2; and a transformed Beta(1,1) for the hyperparameter ρ. The PDF of this Beta(1,1) was,

$$\pi(\rho;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{(\rho_{max}-\rho_{min})\Gamma(\alpha)\Gamma(\beta)}\left(\frac{\rho-\rho_{min}}{\rho_{max}-\rho_{min}}\right)^{\alpha-1}\left(1-\frac{\rho-\rho_{min}}{\rho_{max}-\rho_{min}}\right)^{\beta-1}, \tag{5}$$

where $\rho_{min} = 1/\lambda_{(1)}$ and $\rho_{max} = 1/\lambda_{(n)}$. The prior distributions of the relative proportions of branch lengths to the total tree length and the total tree length were Dirichlet($\underbrace{1,\cdots,1}_{2\cdot 25-3=47}$) and Exponential($10\cdot 47$), respectively.

For all analyses, a single Markov chain was allowed to burn-in for 1,000 cycles, where 1 cycle involved updating all parameters at least once. These updates were effected either by the slice sampling [16] or the Metropolis-Hastings proposal [17, 18]. Following the burn-in period, the chain was allowed to run for 10,000 additional cycles and was sampled once per cycle, providing 10,000 samples. To be practical, the calculation of the likelihood of the general codon model required using a 240-core Nvidia GPU Tesla card for the purpose of high performance computing (HPC).

2.3 Model Selection and Assessment

Two approaches to the model selection and assessment were exploited, namely, the deviance information criterion (DIC) [19] and the conditional predictive ordinate (CPO) [20, 21]. The DIC that reflects both the goodness-of-fit of the model and the complexity of the model is defined as

$$\text{DIC} = \overline{D(\theta)} + p_D \tag{6}$$

$$= -4E_\theta[\log f(\mathbf{y}|\theta)] + 2\log f(\mathbf{y}|\tilde{\theta}), \tag{7}$$

where $\overline{D(\theta)}$ is the posterior mean deviance that shows the goodness-of-fit of the model, p_D is the effective dimension that measures the effective number of

parameters in the model, $\mathbf{y}$ is the data, θ is the model parameters, $\tilde{\theta}$ is the posterior mean of θ, and $f(\mathbf{y}|\theta)$ is the likelihood function that relies on the model. In general, models with smaller DICs should be preferred to those with larger DICs, and increasing p_D improves the goodness-of-fit of the model.

The CPO for the ith observation y_i is defined as,

$$\mathrm{CPO}_i = f\left(y_i|\mathbf{y}_{(-i)}\right) \tag{8}$$

$$= \left(\int \frac{1}{f(y_i|\theta)}\pi(\theta|\mathbf{y})d\theta\right)^{-1} \tag{9}$$

$$\approx \left(\frac{1}{L}\sum_{l=1}^{L}\frac{1}{f(y_i|\theta_l)}\right)^{-1}, \tag{10}$$

where $\mathbf{y}_{(-i)}$ denotes all the observations excluding the ith one, θ is the model parameter vector, $f(y_i|\theta)$ is the likelihood function of the model evaluated at the ith observation, and $\pi(\theta|\mathbf{y})$ is the posterior density. Equation (10) shows the Monte Carlo approximation of CPO_i, where L is the total sample size of an MCMC run, and $f(y_i|\theta_l)$ is the site likelihood of the ith observation evaluated at the lth iteration. It is manifest that CPO_i is calculated without actually deleting the observation y_i from the estimation and is the harmonic mean of the site likelihood of y_i. The CPO statistic is useful to single out outliers or influential observations, with a large value indicating a good fit of the model to the observation y_i and vice versa. A useful summary statistic of the CPO_is is the sum of their logarithms that is called the log pseudo-marginal likelihood (LPML) [20] and defined as,

$$\mathrm{LPML} = \sum_{i=1}^{n}\log(\mathrm{CPO}_i). \tag{11}$$

Models with larger LPMLs have better fit to the data.

3 Results and Discussion

The CAR model is broadly used in spatial statistical studies, for example, demography, epidemiology, biogeography [22], and even single nucleotide polymorphism (SNP) studies [23]. We presented here the first application of the CAR model to the field of phylogenetics for the purpose of detecting positive selection in protein-coding genes. By analyzing the 25 abalone sperm lysin sequences with four spatial correlation schemes, "Simple," "Radius 5," "Radius 10," and "Radius 20," we found the "Radius 10" was the best one among the four candidates using the DIC and LPML methods (Fig. 2). The "Radius 10" scheme obtained the best goodness-of-fit, smallest model dimension and lowest DIC among the four schemes (Fig. 2ABC), and the LPML reinforced the belief (Fig. 2D). The posterior probability of a site

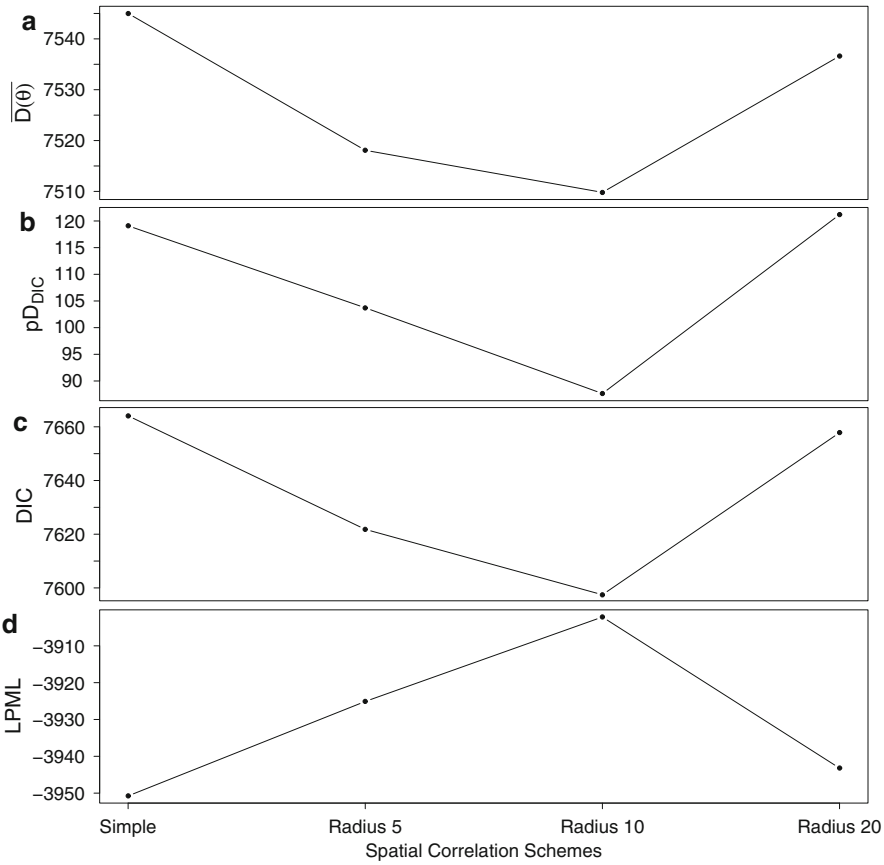


Fig. 2 Assessment of the four different spatial correlation schemes: "Simple," "Radius 5," "Radius 10," and "Radius 20" using the DIC and LPML. (A) The "Radius 10" scheme had the best goodness-of-fit among the 4 schemes; (B) The "Radius 10" scheme had the lowest model dimension; (C) DIC favored the "Radius 10" scheme in all those four; (D) The "Radius 10" scheme had the best LPML among the four.

being under positive selection was reported based on the "Radius 10" scheme (Fig. 3). This quantity was calculated per site from the MCMC output as the fraction of samples whose ω was greater than 1. In the light of the "Radius 10" scheme, sites 1, 3, 5, 9, 16, 24, 27, 31, 32, 35, 40, 58, 59, 66, 70, 78, 81, 97, 104, 111, 114, 117, and 118 were under positive selection with 95 % posterior probability, which is in accordance with the investigation of [12] and [13] (Fig. 3). Those sites were highlighted in black in the cartoon-style structure of the abalone sperm lysin (Fig. 4).

The merits of the CAR model are that: (1) it is a pure Bayesian method; (2) it is compatible with most existing Bayesian phylogenetic software because it does not require changing the likelihood function, making it easy to implement; (3) the model

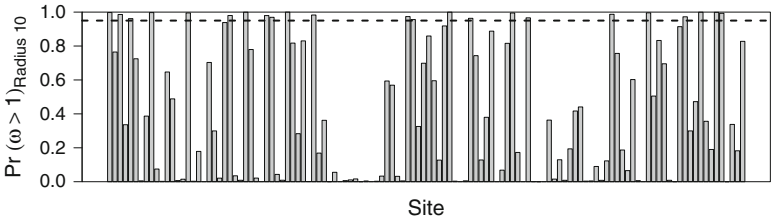


Fig. 3 The posterior probabilities of the abalone sperm lysin data being under positive selection according to the best spatial correlation scheme, "Radius 10." The dotted line shows the 95 % posterior probability cutoff. According to the "Radius 10" scheme, sites 1, 3, 5, 9, 16, 24, 27, 31, 32, 35, 40, 58, 59, 66, 70, 78, 81, 97, 104, 111, 114, 117, and 118 are under the positive selection with 95 % posterior probability.

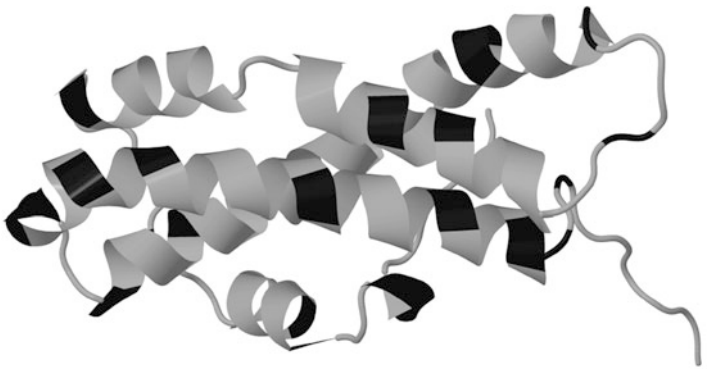


Fig. 4 Image showing the abalone sperm lysin based on the 2LIS PDB data file using Jmol-12.0.49. The sites that are potentially under positive selection are colored black.

itself is quite flexible, and information from other sources can be simply transformed to specify a priori knowledge of the spatial correlation of a protein-coding gene; and (4) the MCMC reaches equilibrium quickly. There are few available choices for detecting positive selection in protein-coding genes at present, and of the available Bayesian frameworks the nonparametric DP model is the most popular one. In general, the CAR model outperformed the DP model in most of the simulation analyses (results not shown): (1) when there was no correlation structure in the simulated data, the CAR model provided better goodness-of-fit (smaller $\overline{D(\theta)}$) than the DP model at the cost of higher model dimension (larger p_D); (2) when the data was simulated from a CAR prior, the CAR model had a clear advantage (superior goodness-of-fit and small model dimension) over the DP model no matter which model selection method was used; and (3) when the data was simulated from a DP prior, the DP model performed better than the CAR model although the advantage of the DP model over the CAR model was small in most analyses.

The LPML quantity based on the CPO statistic is very useful for model selection, particularly considering the simpleness of its calculation. The consistency of LPML

to identify the true model was shown in the simulation study (results not shown). The only drawback of the LPML method is that the site log-likelihoods must be stored during an MCMC run, which is not a common option of most phylogenetic software currently.

The calculation speed of the codon model likelihood is improved greatly because of the usage of a 240-core Nvidia GPU Tesla card and the Beagle-lib [24], a library for evaluating phylogenetic likelihoods by making use of highly parallel processors. The analysis of one spatial correlation scheme requires only 12 hours whereas formerly 7 days are needed to finish 10,000 MCMC cycles. However, in order to bring the graphical card into full play, the likelihood function has to be rewritten, and different graphical card manufacturers have their own standards and languages, which results in limited portability of software.

In this study, we fixed the topology for convenience. There is nothing, however, to prevent the CAR model usage when the topology is allowed to vary during an MCMC run except that more time is needed, particularly when codon models are used.

References

[1] Yang, Z.: A space-time process model for the evolution of DNA sequences. Genetics **139**, 993–1005 (1995)
[2] Felsenstein, J. and G. A. Churchill: A Hidden Markov Model approach to variation among sites in rate of evolution. Molecular Biology and Evolution **13**, 93–104 (1996)
[3] Koshi, J. M. and R. A. Goldstein: Models of natural mutations including site heterogeneity. Proteins **32**, 289–295 (1998)
[4] Liò, P., N. Goldman, J. L. Thorne, and D. T. Jones3: PASSML: combining evolutionary inference and protein secondary structure prediction. Bioinformatics **14**, 726–733 (1998)
[5] Liò, P. and N. Goldman: Using protein structural information in evolutionary inference: transmembrane proteins. Molecular Biology and Evolution **16**, 1696–1710 (1999)
[6] Robinson, D., D. Jones, H. Kishino, N. Goldman, and J. Thorne: Protein evolution with dependence among codons due to tertiary structure. Molecular Biology and Evolution **20**, 1692–1704 (2003)
[7] Rodrigue, N., N. Lartillot, D. Bryant, and H. Philippe: Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene **347**, 207–217 (2005)
[8] Kleinman, C. L., N. Rodrigue, N. Lartillot, and H. Philippe: Statistical potentials for improved structurally constrained evolutionary models. Molecular Biology and Evolution **27**, 1546–1560 (2010)
[9] Huelsenbeck, J., S. Jain, S. Frost, and S. Pond: A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proceedings of the National Academy of Sciences of the United States of America **103**, 6263–6268 (2006)
[10] Besag, J.: Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society. Series B (Methodological) **36**, 192–236 (1974)
[11] Banerjee, S., B. P. Carlin, and A. E. Gelfand: Hierarchical modeling and analysis for spatial data. Chapman & Hall/CRC, London (2004)
[12] Yang, Z., W. Swanson, and V. Vacquier: Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Molecular Biology and Evolution **17**, 1446–1455 (2000)

[13] Yang, Z. and J. Bielawski: Statistical methods for detecting molecular adaptation. Trends in Ecology and Evolution **15**, 496–503 (2000)
[14] Berman, H., K. Henrick, and H. Nakamura: Announcing the worldwide Protein Data Bank. Nature Structural Biology **10**, 980–980 (2003)
[15] Kresge, N., V. D. Vacquier, and C. D. Stout: 1.35 and 2.07 A resolution structures of the red abalone sperm lysin monomer and dimer reveal features involved in receptor binding. Acta Crystallographica Section D: Biological Crystallography **56**, 34–41 (2000)
[16] Neal, R. M.: Slice sampling. Annals of Statistics **31**, 705–741 (2003)
[17] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller: Equation of state calculations by fast computing machines. The Journal of Chemical Physics **21**, 1087–1092 (1953)
[18] Hastings, W. K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
[19] Spiegelhalter, D., N. Best, B. Carlin, and A. Linde: Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society. Series B (Statistical Methodology) **64**, 583–639 (2002)
[20] Geisser, S. and W. F. Eddy: A predictive approach to model selection. Journal of the American Statistical Association **74**, 153–160 (1979)
[21] Chen, M.-H., Q.-M. Shao, and J. G. Ibrahim: Monte Carlo methods in Bayesian computation. Springer-Verlag Inc., Berlin, New York (2000)
[22] Gelfand, A. E., J. A. Silander, S. Wu, A. Latimer, P. O. Lewis, A. G. Rebelo, and M. Holder: Explaining species distribution patterns through hierarchical modeling. Bayesian Analysis **1**, 41–91 (2006)
[23] Guo, F., D. K. Dey, and K. E. Holsinger: A Bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. Journal of the American Statistical Association **104**, 142–154 (2009)
[24] Suchard, M. A. and A. Rambaut: Many-core algorithms for statistical phylogenetics. Bioinformatics **25**, 1370–1376 (2009)

Dimension Reduction for Tensor Classification

Peng Zeng and Wenxuan Zhong

Abstract This article develops a sufficient dimension reduction method for high dimensional regression with tensor predictors, which extends the conventional vector-based dimension reduction model. It proposes a tensor dimension reduction model that assumes that a response depends on some low-dimensional representation of tensor predictors through an unspecified link function. A sequential iterative dimension reduction algorithm (SIDRA) that effectively utilizes the tensor structure is proposed to estimate the parameters. The SIDRA generalizes the method in Zhong and Suslick (2012), which proposes an iterative estimation algorithm for matrix classification. Preliminary studies demonstrate that the tensor dimension reduction model is a rich and flexible framework for high dimensional tensor regression, and SIDRA is a powerful and computationally efficient method.

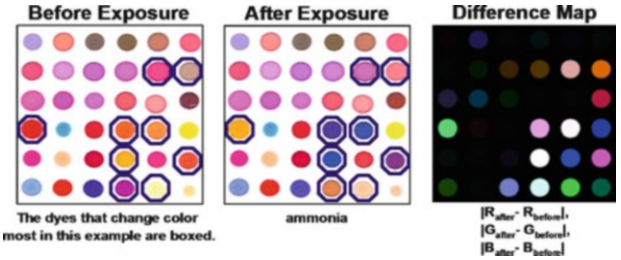
1 Introduction

A tensor is a multidimensional array. More precisely, an *m*th-order tensor is an element of the tensor product of *m* vector spaces, each of which has its own coordinate system termed mode Merris (1997). For example, a first-order tensor is a vector with one mode and a second-order tensor is a matrix with two modes (row and column). With the rapid development of science and technology in the

P. Zeng
Department of Mathematics & Statistics, Auburn University, Auburn, AL 36849, USA
e-mail: zengpen@auburn.edu

W. Zhong (✉)
Department of Statistics, The University of Georgia, Athens,
GA 30602-7952, USA
e-mail: wenxuan@uga.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_18,



	$\lvert R_{after} - R_{before}\rvert$	$\lvert R_{after} - R_{before}\rvert$	$\lvert R_{after} - R_{before}\rvert$
dye1	–0.814	–0.652	–0.264
dye2	–0.806	–0.099	–1.429
…			
dye36	0.991	–1.038	0.823

Fig. 1 The difference map of a colorimetric array can be generated for any odorant or mixture of odorants by digital subtraction, pixel by pixel, of the image of the array before and after exposure.

past decades, large amount of tensor observations is routinely collected, processed, and stored in many scientific researches and commercial activities nowadays. The colorimetric sensor array (CSA) data is one of such examples.

CSA, invented by Kenneth Suslick's group (UIUC), is one of the leading electronic nose techniques. As shown in Fig. 1, CSA is simply a digitally-imaged, two-dimensional extension of litmus paper (Rakow et al. 2005; Rakow and Suslick 2000; Zhang and Suslick 2005). For any odorant, a response is generated by digital subtraction, pixel by pixel, of the color of p pre-print chemo-responsive dyes before and after exposure: red value after exposure (R_{after}) minus red value before (R_{before}), green (G_{after}) minus green (G_{before}), blue (B_{after}), minus blue (B_{before}). Averaging the centers of the spots (~ 300 pixels) for each dye, the result is simply a $p \times 3$ matrix (second-order tensor), where each row represents the color change of a dye and each column represents one of the three spectrum coordinates (R,G,B) of a color cube. If we further consider the time and the odorant concentration as factors, we will have forth-order tensor (dyes $\times$ color spectrum $\times$ time $\times$ odorant concentration). The tensor is a color fingerprint for each odorant and will be used for downstream statistical analysis. For example, we can use CSA to discriminate an unknown odorant (Feng et al. 2010; Lim et al. 2008; Zhang et al. 2006), evaluate a wine's taste (Savage 2012), monitor bacteria growth (Carey et al. 2011) and even cancer risk prediction (Ehmann et al. 2012; Morgan 2012).

Existing statistical methods for the tensor data largely ignore the tensor structure by simply vectorizing each tensor observation into a vector and offering solutions using the vector-based statistical methods. These solutions, however, are far from satisfactory. First, the simple vectorization destroys the original design information and leads to interpretation difficulties. For example, a tensor presentation of the CSA data can help us answer the question such as which dye is classification-relevent on what color spectrum. An answer to the previous question cannot be given if we simply stack the columns of the observation and use a vector-based classification method because we have no separate parameters for the dyes and the color spectrums. Second, the simple vectorization significantly aggravates the curse of dimensionality, which refers to various difficulties that a large number of variables (or dimensions) can cause to function approximation, model fitting,

information extraction, and computation (Fan and Li 2006). For example, typical 36-dye CSA data require $36 + 3$ parameters, while simple vectorization generates $36 \times 3 = 108$ parameters for a conventional classification analysis. Thus the simple vectorization renders many vector-based approaches infeasible for a sample with less than 108 observations. Moreover, even if we have fairly large sample size, both the computational efficiency and the estimation accuracy of the classical vector-based analysis will be compromised by simple vectorization (Donoho and Elad 2003; Fan and Li 2006). There is an obvious pressing need for new statistical methods and theories that can directly utilize the intrinsic tensor structures.

Regression analysis is probably the most popular statistical tool for modeling the relationship between a response Y and a series of predictor variables X. With $X \in \mathbb{R}^p$, various models and methods have been developed for regression analysis in the literature, ranging from classic linear regression to nonparametric regression. In general, regression can be considered as an inference about the conditional distribution of Y given X, often with the mean response $E(Y \mid X)$ of particular interest. When p is large, the estimate of $E(Y \mid X)$ can be seriously compromised by the curse of dimensionality. To overcome this limitation, a wide range of methods have been proposed to facilitate dimension reduction in regression in the literature. Sufficient dimension regression (SDR) is one of such proposals, which assumes that the response Y only depends on a lower dimensional projection of the X. To identify the lower dimensional projections is the major goal of SDR analysis.

Although the SDR model is arguably the most general model that can be used in a wide spectrum of applications, it is not directly applicable to the regression model with tensor predictors, because the tensor structure will be lost in the lower dimensional projection if we simply vectorize the tensor predictor. To overcome this limitation, Li et al. (2010) proposed dimension folding method to obtain the lower dimensional projection in the form of a tensor product. Briefly, Li et al. (2010) find the lower dimensional projections for the rows and columns of the matrix predictor separately such that their tensor product contains the smallest vector-based lower dimensional projection.

Although the dimension folding method is promising in keeping the tensor structure during dimension reduction, we always wonder if there exists a space smaller than the space targeted by dimension folding and also containing all the regression-related lower dimensional projection of X. Another related question is how to find the smaller space if there exists one. In this article, we show that a smaller space indeed exists. To recover the space, we propose a tensor dimension reduction (TDR) model. A sequential iterative dimension reduction algorithm (SIDRA) is proposed to estimate the parameters in the TDR model. The SIDRA generalizes the recent work of Zhong and Suslick (2012) on matrix classification to the dimension reduction regression model under the sufficient dimension reduction framework.

The rest of the article is organized as follows. In Sect. 2, we give a brief introduction to sufficient dimension reduction regression (DR) model that is proposed by Li (1991). The DR model is generalized to the tensor dimension reduction regression

model (TDR) in Sect. 3 to allow tensor predictors. Section 4 discusses the population version of the SIDRA procedure. Several implementation issues of the procedure are discussed in Sect. 5. Some case studies are reported in Sect. 6. Additional remarks in Sect. 7 conclude the article.

2 Sufficient Dimension Reduction

Let $X \in \mathbb{R}^p$ be a random vector and $Y \in \mathbb{R}$ be a random scalar. Suppose $\mathcal{S}$ is a subspace of $\mathbb{R}^p$ and $P_{\mathcal{S}}$ is the projection operator from $\mathbb{R}^p$ to $\mathcal{S}$ in the standard inner product. If

$$Y \perp X | P_{\mathcal{S}} X, \tag{1}$$

where $\perp$ means statistical independence, then it is said that $P_{\mathcal{S}}X$ is sufficient for the dependence of Y on X. In other words, the projection $P_{\mathcal{S}}X$ captures all the information contained in X regarding Y. Model (1) was formally proposed by Cook and Weisberg (1994) and was further discussed by Cook (1996) and Cook (1998). Model (1) is equivalent to several other formulations, for example, the general multiple-index model

$$Y = h(\beta_1^T X, \cdots, \beta_d^T X, \varepsilon) \tag{2}$$

proposed in Li (1991), where h is an unknown function, β_is are p-dimensional vectors of unit length, d is an integer less than p, T denotes transpose, ε is independent of X, and $\mathrm{E}(\varepsilon) = 0$. Given $(\beta_1^T X, \cdots, \beta_d^T X)$, Y and X are independent. Therefore the linear space spanned by β_i can serve as $\mathcal{S}$ in model (1). Conversely, if (1) holds, there exist h and ε such that (2) holds. A brief proof of the equivalence between the two models can be found in Zeng and Zhu (2010).

Model (1) and (2) are referred to as the dimension reduction regression (DR) model and $\mathcal{S}$ is referred to as a dimension reduction subspace. Dimension reduction subspace may not be unique. Cook (1996) introduced an important concept called *central space*, which is defined as the intersection of all dimension reduction subspaces when it is a dimension reduction subspace itself. The central subspace is denoted by $\mathcal{S}_{Y|X}$, and the dimension of $\mathcal{S}_{Y|X}$ is called the structural dimension of regressing Y on X. Under mild conditions, it can be shown that $\mathcal{S}_{Y|X}$ exists; see Cook and Weisberg (1994) for details. Throughout this article, we assume the existence of $\mathcal{S}_{Y|X}$.

The dimension reduction regression model is a very general formulation and covers a wide range of parametric and semi-parametric models. For example, if Y is a discrete variable taking values in $\{1, 2, \cdots, K\}$, the dimension reduction regression model covers logistic regression and many classification models. If Y is a continuous variable taking values in $\mathbb{R}$, linear regression model, partial linear model, and single index model are its special cases.

It was showed in Chen and Li (1998) and Zhong et al. (2012) that, when X satisfies the so-called linearity condition $\mathrm{E}(X|P_{\mathcal{S}_{Y|X}}X) = P_{\mathcal{S}_{Y|X}}X$, a basis of $\mathcal{S}_{Y|X}$, i.e. $\{\beta_1, \cdots, \beta_d\}$ can be obtained by recursively maximizing

$$L(\eta) = \max_{T} \mathrm{corr}^2(T(Y), \eta^T X) \tag{3}$$

subject to the constraint that $\beta_i \Sigma_X \beta_j = \delta_{ij}$, where $\Sigma_X \overset{\triangle}{=} \mathrm{var}(X)$ and $\delta_{ij} = 1$ if $i = j$ and $= 0$ otherwise, $\eta \in \mathbb{R}^p$ and T is any possible transformations of Y including non-monotone ones. In general, the $L(\eta)$ reflects the largest possible squared correlation between a transformed response $T(Y)$ and the projection $\eta^T X$. At the population level, Chen and Li (1998) showed that $L(\eta)$ has an explicit form

$$L(\eta) = \frac{\eta^T \mathrm{var}[\mathrm{E}(X|Y)]\eta}{\eta^T \mathrm{var}(X)\eta} = \frac{\eta^T M \eta}{\eta^T \Sigma_X \eta}, \tag{4}$$

where $M \overset{\triangle}{=} \mathrm{var}[\mathrm{E}(X|Y)]$. Therefore, $\beta_1, \cdots, \beta_d$ are the eigenvectors of M with respect to Σ_X corresponding to the largest d eigenvalues.

Based on the definition of M, a computationally stable and fast procedure called sliced inverse regression (SIR) was proposed in Li (1991) to generate an estimate of M, which is further used to generate an estimate of $P_{\mathcal{S}_{Y|X}}$. Observing (X_i, Y_i), the SIR algorithm is described as follows: (i) Divide the range of Y into several disjoint intervals $I_1, \cdots, I_H$ and let n_h denote the number of observations falling in I_h; (ii) Estimate $\mathrm{E}(X)$ and Σ_X by the sample mean $\bar{X}$ and sample variance covariance matrix $\hat{\Sigma}_X$ of X and estimate $\mathrm{E}(X \mid Y)$ by $\bar{X}_h = \frac{1}{n_h} \sum_{\{i: Y_i \in I_h\}} X_i$; (iii) Estimate M by $\hat{M}$, where $\hat{M} = \frac{1}{n} \sum_{h=1}^{H} n_h (\bar{X}_h - \bar{X})(\bar{X}_h - \bar{X})^T$. (iv) Apply the spectral decomposition to $\hat{M}$ with respect to $\hat{\Sigma}_X$ to obtain its eigenvalue-eigenvector pairs $(\hat{\lambda}_i, \hat{\beta}_i)$ where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$. The $\hat{\beta}_i$ is the SIR estimate of β_i and is referred to as the ith SIR direction. Empirical studies demonstrate that SIR is fairly successful in achieving dimension reduction for the high dimensional regression. The properties and asymptotic behaviors of SIR have been well studied in the literature; see Duan and Li (1991) and Hsing and Carroll (1992), among others.

3 Tensor Sufficient Dimension Reduction

The dimension reduction regression model discussed in the preceding section can be extended to second or higher order tensors. To ease the presentation, this section only focuses on second order tensors. The model and algorithm proposed here can be similarly extended to mth-order tensors.

We introduce some notations and conventions used in the presentation. Denote $\alpha \otimes \beta$ as the Kronecker product of vectors α and β. For two linear spaces, $\mathcal{S}_1 \subset \mathbb{R}^{p_1}$ and $\mathcal{S}_2 \subset \mathbb{R}^{p_2}$, denote $\mathcal{S}_1 \otimes \mathcal{S}_2$ as the linear space spanned by $\{\alpha \otimes \beta \mid \alpha \in \mathcal{S}_1, \beta \in \mathcal{S}_2\}$. With a little bit abuse of notation, for any second-order tensor $\mathbf{X} \in \mathbb{R}^{p_1} \otimes \mathbb{R}^{p_2}$, we let $X \in \mathbb{R}^{p_1 p_2}$ be the vector obtained by stacking up the columns of $\mathbf{X}$.

3.1 Decomposable Tensor and Core Tensor Dimension Reduction Set

Assume that the response $Y \in \mathbb{R}$ depends on a low-dimensional projection of X through an unknown link function g, i.e.,

$$Y = g(\gamma_1^T X, \ldots, \gamma_d^T X, \varepsilon), \tag{5}$$

where g is an unknown function, $\gamma_i = \beta_i \otimes \alpha_i$ with $\alpha_i \in \mathbb{R}^{p_1}$ and $\beta_i \in \mathbb{R}^{p_2}$ are indexes, and ε is a random error independent of X. Model (5) is equivalent to say that Y is independent of X given $(\gamma_1^T X, \ldots, \gamma_d^T X)$. Note that $\gamma_1, \ldots, \gamma_d$ may not be identifiable because γ_i for $i = 1, \ldots, d$ and g are confounded with each other. However, the space spanned by $\gamma_1, \ldots, \gamma_d$ is identifiable. For example, consider model $Y = g(\alpha_1^T \mathbf{X} \beta_1, \alpha_2^T \mathbf{X} \beta_1) + \varepsilon$. For this model, we can think the indexes are $\beta_1 \otimes \alpha_1$ and $\beta_1 \otimes \alpha_2$ and can also think the indexes are $\beta_1 \otimes (\alpha_1 + \alpha_2)$ and $\beta_1 \otimes (\alpha_1 - \alpha_2)$. But in either case, they span the same linear space.

Define $\mathcal{S}$ as the space spanned by $(\gamma_1, \cdots, \gamma_d)$, $\mathcal{S}_1$ as the space spanned by $(\alpha_1, \cdots, \alpha_d)$ and $\mathcal{S}_2$ as the space spanned by $(\beta_1, \cdots, \beta_d)$. It is worth noting that not all the vectors in $\mathcal{S}$ can be written as a Kronecker product of a α and a β, where $\alpha \in \mathcal{S}_1$ and $\beta \in \mathcal{S}_2$. For example, for $\beta_1 \otimes \alpha_1 \in \mathcal{S}$ and $\beta_2 \otimes \alpha_2 \in \mathcal{S}$, we have $\beta_1 \otimes \alpha_1 + \beta_2 \otimes \alpha_2 \in \mathcal{S}$. But, in general, it cannot be expressed as a Kronecker product of a $\alpha \in \mathcal{S}_1$ and a $\beta \in \mathcal{S}_2$ unless $\alpha_1 = \alpha_2$ or $\beta_1 = \beta_2$. To keep the tensor structure and the interpretability of the indexes in model (5), we introduce the concept of decomposable tensor. An mth order tensor $\gamma \in \mathcal{S}_1 \otimes \mathcal{S}_2 \otimes \cdots \otimes \mathcal{S}_m$ is said to be decomposable if it can be written in the form of $\eta_1 \otimes \eta_2 \otimes \cdots \otimes \eta_m$ where $\eta_i \in \mathcal{S}_i$ for $i = 1, \ldots, m$ (Merris 1997). Thus, for model (5), we let $\mathbb{S}$ be the set that consists of all the decomposable tensors in $\mathcal{S}$. It is easy to see that $\mathbb{S}$ is identifiable if $\mathcal{S}$ is identifiable. From now on, we define $\mathbb{S}$ as the tensor dimension reduction set (TDRS). Similar to the dimension reduction subspace discussed in Sect. 2, the TDRS might not be unique. To overcome this challenge, we introduce the core tensor dimension reduction set (CTDRS), which is defined as the intersection of all TDRSs when it is a TDRSs itself. The rank of the CTDRS is referred to as the core dimension of regressing Y on X.

It is important to point out that model (5) is substantially different from their counterpart in the preceding section although they look similar in expression. For simplicity, consider the case when $d = 1$. If we ignore the fact that $\mathbf{X}$ is a second-order tensor and treat it as a vector, the index β in model (2) has $p_1 p_2$ parameters. However, if we treat it as a second-order tensor, noting that $\gamma_1 = \beta_1 \otimes \alpha_1$, the index γ_1 has only $p_1 + p_2$ parameters. Intuitively, considering the tensor structure in estimation will yield more efficient estimates due to less number of parameters. For the central subspace in (1), the basis can be any set of d linearly independent vectors, while in model (5), the index is specified as a set of d linearly independent vectors of the form $\{\beta_i \otimes \alpha_i, i = 1, \ldots, d\}$.

3.2 Comparison with Dimension Folding

In Li et al. (2010), a different tensor dimension reduction model called dimension folding is proposed based on the following model,

$$Y \perp X | (P_{\mathcal{S}_2} \otimes P_{\mathcal{S}_1}) X, \tag{6}$$

where $\mathcal{S}_1$ the linear space spanned by $\{\alpha_1, \ldots, \alpha_d\}$, $\mathcal{S}_2$ is the linear space spanned by $\{\beta_1, \ldots, \beta_d\}$, and $P_{\mathcal{S}_i}$ is a projection operator from $\mathbb{R}^{p_i}$ to $\mathcal{S}_i$ for $i = 1, 2$ in the standard inner product. Because α_i for $i = 1, \cdots, d$ operate on the rows of $\mathbf{X}$, $\mathcal{S}_1$ is named as the left dimension-folding subspace in Li et al. (2010). Similarly, $\mathcal{S}_2$ is named as the right dimension-folding subspace in Li et al. (2010). The smallest left dimension-folding subspace $\mathcal{S}_1^c$ is called the central left dimension-folding subspace for $Y|\mathbf{X}$, and the smallest right dimension-folding subspace $\mathcal{S}_2^c$ is called the central right dimension-folding subspace for $Y|\mathbf{X}$. Li et al. (2010) also defined the central dimension-folding subspace as $\mathcal{S}_2^c \otimes \mathcal{S}_1^c$ and showed that

$$\mathcal{S}_{Y|X} \subset \mathcal{S}_2^c \otimes \mathcal{S}_1^c,$$

where $\mathcal{S}_{Y|X}$ is the central subspace obtained by treating X as a vector and completely ignoring its tensor structure.

It is important to point out that $\mathcal{S}$ in model (5) is substantially different from the central dimension-folding subspace in Li et al. (2010). In fact, model (5) is a more parsimonious model for the relationship between Y and X. Let $\mathcal{S}_{Y|\mathbf{X}}$ be the linear space that spanned by $\mathbb{S}$, the CTDRS. Then, it is easy to show that

$$\mathcal{S}_{Y|X} \subset \mathcal{S}_{Y|\mathbf{X}} \subset \mathcal{S}_2^c \otimes \mathcal{S}_1^c.$$

This claim can be illustrated clearly using the following toy example. Consider model $Y = g(\gamma_1^T X, \gamma_2^T X, \varepsilon)$, where $\gamma_1 = \beta_1 \otimes \alpha_1$ and $\gamma_2 = \beta_2 \otimes \alpha_2$. Then $\mathcal{S}_1^c = \text{span}(\alpha_1, \alpha_2)$ and $\mathcal{S}_2^c = \text{span}(\beta_1, \beta_2)$. Hence $\mathcal{S}_2^c \otimes \mathcal{S}_1^c$ is spanned by $(\beta_1 \otimes \alpha_1, \beta_1 \otimes \alpha_2, \beta_2 \otimes \alpha_1, \beta_2 \otimes \alpha_2)$ with four dimensions, while $\mathcal{S}_{Y|\mathbf{X}}$ is spanned by $(\beta_1 \otimes \alpha_1, \beta_2 \otimes \alpha_2)$ with two dimensions. The latter is clearly a proper subset of the former. This example clearly demonstrates that the space $\mathcal{S}_{Y|\mathbf{X}}$ defined in (5) is a subspace of the central dimension-folding subspace.

4 Tensor Dimension Reduction and SIDRA Procedure

From now on, we focus on the estimation of the γ_i, $1 \le i \le d$, while leaving g unspecified. Let $\mathcal{R}$ be the set of all decomposable tensors in $\mathbb{R}^{p_2} \otimes \mathbb{R}^{p_1}$. Similar to the estimation of model (1) and (2), the first direction γ_1 can be obtained by maximizing $L(\eta)$ in $\mathcal{R}$, i.e.,

$$\gamma_1 = \text{argmax}_{\eta \in \mathcal{R}} \frac{\eta^T \text{var}[\text{E}(X|Y)]\eta}{\eta^T \text{var}(X)\eta}. \tag{7}$$

Comparing (7) to (4), we may naturally attempt the optimization of (7) using the tensor spectrum decomposition as what we did for the vector predictors. However, tensor spectral decomposition, such as the PARAFAC (Harshman and Lundy 1984) or Tucker model (Tucker 1951; 1966), unlike its vector sibling, may not maximize $L(\eta)$. Beyond this, the definition and algorithm of the tensor spectral decomposition are far from mature and have many intrinsic problems. For example, the orthogonality on each mode is not assumed and decomposition on the same mode is not unique using different algorithms (Kolda and Bader 2009; Smilde et al. 2004).

Recall that X is the vectorization of $\mathbf{X}$, where $\mathbf{X} \in \mathbb{R}^{p_2} \otimes \mathbb{R}^{p_1}$, and for any $\eta \in \mathcal{R}$, we have $\eta = \beta \otimes \alpha$ and $\eta^T X = \alpha^T \mathbf{X} \beta$, where $\alpha \in \mathbb{R}^{p_1}$ and $\beta \in \mathbb{R}^{p_2}$. Maximizing (7) is equivalent to maximizing the bivariate function

$$L_1(\alpha, \beta) = \frac{\text{var}[E(\alpha^T \mathbf{X} \beta \mid Y)]}{\text{var}(\alpha^T \mathbf{X} \beta)} \tag{8}$$

with respect to α and β. Practically, the maximizer of (8) can be obtained by iteratively maximizing the following univariate functions

$$L_\alpha(\eta) = \frac{\eta^T \text{var}[\text{E}(\mathbf{X}^T \alpha|Y)]\eta}{\eta^T \text{var}(\mathbf{X}^T \alpha)\eta} \tag{9}$$

for given α and

$$L_\beta(\eta) = \frac{\eta^T \text{var}[\text{E}(\mathbf{X}\beta|Y)]\eta}{\eta^T \text{var}(\mathbf{X}\beta)\eta} \tag{10}$$

for given β.

Notice that Eq. (9) is the unrestricted maximization in the vector space $\mathbb{R}^{p_2}$. Thus, if α is given, β can be obtained by SIR. The maximization of (10) is similar to (9). Equations (9) and (10) imply that α and β can be estimated iteratively by SIR if an initial estimate of α is provided. Notice that the objective function in Eq. (8) is bounded above by one and the iterative maximization approach ensures the increasing of $L(\cdot)$ in each iteration. Thus, the convergence is guaranteed for the iterative algorithm.

Let $\gamma_1 = (\beta_1 \otimes \alpha_1)$ be the maximizer of (8). The second direction γ_2 can be obtained in a similar way by maximizing Eq. (8) on all the decomposable tensors of $\mathbb{R}^{p_2} \otimes \mathbb{R}^{p_1}$ that is orthogonal to γ_1, where the orthogonality is defined using the standard inner product, i.e., $\gamma_2^T \gamma_1 = 0$. Thus, we can obtain directions in a sequential fashion. Generally, in the $(k+1)$th step, we project the tensor predictors onto the space that is orthogonal to the space spanned by the previously obtained $(\gamma_1, \cdots, \gamma_k)$ and find a new γ_{k+1} by iteratively maximizing (8) with X replaced by the projection. More precisely, given $M_k = (\gamma_1, \cdots, \gamma_k)$, define $P_k = \Sigma_X M_k (M_k^T \Sigma_X M_k)^{-1} M_k^T$ as the

projection matrix from $\mathbb{R}^{p_1 p_2}$ onto the space spanned by M_k with respect to Σ_X. Let $X_{(k)} \in \mathbb{R}^{p_1 p_2}$ be the projection of X in the complementary space of the linear space that spanned by M_k, i.e.,

$$X_{(k)} = (I - P_k)X. \tag{11}$$

Let $\mathcal{R}_k$ denote the set of all decomposable tensors of $\mathbb{R}^{p_2} \otimes \mathbb{R}^{p_1}$ that is orthogonal to the space spanned by $(\gamma_1, \cdots, \gamma_k)$. Then we have

$$\gamma_{k+1} = \text{argmax}_{\eta \in \mathcal{R}_k} \frac{\eta^T \text{var}[\text{E}(X_{(k)}|Y)]\eta}{\eta^T \text{var}(X_{(k)})\eta}. \tag{12}$$

Replace $X_{(k)}$ in (12) by the matrization of X_k, denote by $\mathbf{X}_{(k)}$. The objective function in (12) can be rewritten as

$$L_{k+1}(\alpha, \beta) = \frac{\text{var}[E(\alpha^T \mathbf{X}_{(k)} \beta \mid Y)]}{\text{var}(\alpha^T \mathbf{X}_{(k)} \beta)}. \tag{13}$$

Thus, we can obtain the $(k+1)$th direction of the CTDRS using the iterative algorithm that we proposed for estimating γ_1. Our algorithm for estimating $(\gamma_1, \cdots, \gamma_d)$ can be summarized as follows.

Algorithm 1 The sequential iterative dimension reduction algorithm (SIDRA)

- Let $\mathbf{X}_{(0)} = \mathbf{X}$ and M_0 be an empty set.
- Do $k = 1, 2, \ldots$ for the following and stop when $L_k(\alpha_k^*, \beta_k^*) = 0$.
  - Choose an initial value of α_k, for example, a random vector of unit length.
  - Iterate the following until converge to find α_k^* and β_k^* that maximize (13).
    - Update β_k^* by maximizing $L_{\alpha_k^*}(\eta)$ with $\mathbf{X}$ replaced by $\mathbf{X}_{(k-1)}$.
    - Update α_k^* by maximizing $L_{\beta_k^*}(\eta)$ with $\mathbf{X}$ replaced by $\mathbf{X}_{(k-1)}$.
    - Stop iteration when the difference between two consecutive $L_k(\alpha_k^*, \beta_k^*)$ is 0.
  - Let $\gamma_k = \beta_k^* \otimes \alpha_k^*$ and calculate $L_k(\alpha_k^*, \beta_k^*)$.
  - Let $M_k = M_{k-1} \cup \{\gamma_k\}$ if $L_k(\alpha_k^*, \beta_k^*) > 0$ and calculate $\mathbf{X}_{(k)}$ as in (11).
- Output $\gamma_1, \ldots, \gamma_k$, which are in M_k.

5 Implementation

With observations $(\mathbf{X}_i, Y_i)$ for $i = 1, \cdots, n$, the maximization steps in Algorithm 1 can be accomplished using SIR as described in Sect. 2. More precisely, we follow the steps below to estimate d directions, $\hat{\gamma}_1, \ldots, \hat{\gamma}_d$, which are referred to as the tensor SIR directions. Note that the number of directions d should be smaller than the number of slices H.

- Let $\mathbf{X}_{i,(0)} = \mathbf{X}_i$ for $i = 1, \ldots, n$. Calculate $\hat{\Sigma}_X$ as the covariance matrix of X_i, the vecterization of $\mathbf{X}_i$.
- Do $k = 1, 2, \ldots, d$ for the following.
  - Choose an initial value of α_k, for example, a random vector of unit length.
  - Iterate the following until converge to find $\hat{\alpha}_k$ and $\hat{\beta}_k$.
    - Find $\hat{\beta}_k$ as the first SIR direction with y_i as response and $\mathbf{X}_{i,(k)}^T \hat{\alpha}_k$ as predictor.
    - Find $\hat{\alpha}_k$ as the first SIR direction with y_i as response and $\mathbf{X}_{i,(k)} \hat{\beta}_k$ as predictor.
  - Let $\hat{\gamma}_k = \hat{\beta}_k \otimes \hat{\alpha}_k$ and $\hat{M}_k = (\hat{\gamma}_1, \ldots, \hat{\gamma}_k)$.
  - Calculate $X_{i,(k)} = (I - \hat{P}_k) X_i$, where $\hat{P}_k = \hat{\Sigma}_X \hat{M}_k (\hat{M}_k^T \hat{\Sigma}_X \hat{M}_k)^{-1} \hat{M}_k^T$.
- Output $\hat{\gamma}_1, \ldots, \hat{\gamma}_k$ as the first d tensor SIR directions.

Although the SIDRA method can effectively recover the core tensor dimension reduction set theoretically and intuitively, we may encounter some practical problem during implementation. For example, if the true model is $Y = g((\beta_1 \otimes \alpha_1)^T X, (\beta_2 \otimes \alpha_1)^T X, \varepsilon)$, the $\hat{\alpha}_1$ obtained in the first step may not be the same as the $\hat{\alpha}_1$ obtained in the second step empirically though they will be in general very close. To face this dilemma, the SIDRA algorithm need to be modified correspondingly so that the $\hat{\alpha}$'s and $\hat{\beta}$'s obtained in different iterations are either identical or orthogonal to each other.

In the following, we present the modified algorithm more precisely. Let $\mathcal{S}_{1,k}$ and $\mathcal{S}_{2,k}$ be the space spanned by $(\hat{\alpha}_1, \ldots, \hat{\alpha}_k)$ and $(\hat{\beta}_1, \ldots, \hat{\beta}_k)$, respectively. It is clear that assuming $\hat{\gamma}_{k+1}^T \hat{\gamma}_i = 0$ is equivalent to assuming $\hat{\beta}_{k+1}^T \hat{\beta}_i = 0$ or $\hat{\alpha}_{k+1}^T \hat{\alpha}_i = 0$, because $\hat{\gamma}_{k+1}^T \hat{\gamma}_i = 0 = (\hat{\beta}_{k+1}^T \hat{\beta}_i)(\hat{\alpha}_{k+1}^T \hat{\alpha}_i)$ for $i = 1, \ldots, k$. Therefore, to obtain $\hat{\gamma}_{k+1}$ that is orthogonal to $\hat{\gamma}_i$ for $i = 1, \ldots, k$, we only need to find the maximizer of (13) in the following 4 subsets

$$D_1 = \{(\alpha, \beta) | \alpha \in \mathcal{S}_{1,k}^{\perp}, \beta \in \mathcal{S}_{2,k}\},$$

$$D_2 = \{(\alpha, \beta) | \alpha \in \mathcal{S}_{1,k}, \beta \in \mathcal{S}_{2,k}^{\perp}\},$$

$$D_3 = \{(\alpha, \beta) | \alpha \in \mathcal{S}_{1,k}^{\perp}, \beta \in \mathcal{S}_{2,k}^{\perp}\},$$

$$D_4 = \{(\alpha, \beta) | \alpha \in \mathcal{S}_{1,k}, \beta \in \mathcal{S}_{2,k}\},$$

separately and then select the one with the largest value of $L_k(\hat{\alpha}, \hat{\beta})$, where $\mathcal{S}_{i,k}^{\perp}$ for $i = 1, 2$ is referred to as the complementary space of $\mathcal{S}_{i,k}$. Clearly, the newly identified directions only introduce the direction that is orthogonal to the existing directions by this way.

To make our presentation concise, we use D_4 as an example to illustrate how to find the maximizer of (13) on that domain. Let $A = (\hat{\alpha}_1, \cdots, \hat{\alpha}_k)$ be a set of basis of

$\mathcal{S}_{1,k}$. Consequently, $\alpha = Aa$ for some vector a, if $\alpha \in \mathcal{S}_{1,k}$. Similarly, we can write $\beta = Bb$ for some vector b, where $B = (\hat{\beta}_1, \cdots, \hat{\beta}_k)$ is a basis of $\mathcal{S}_{2,k}$. Therefore, the objective function (13) can be rewritten as

$$\tilde{L}_k(a,b) = \frac{\text{var}[\text{E}(a^T A^T X_{(k)} Bb \mid Y)]}{\text{var}(a^T A^T X_{(k)} Bb)}. \tag{14}$$

Define $\tilde{X}_{(k)} = A^T X_{(k)} B$. Then we can use the iterative algorithm proposed in Algorithm 1 to maximize (14) with respect to a and b. Denote the maximizer as $\hat{a}$ and $\hat{b}$. Finally, we can calculate $\hat{\alpha} = A\hat{a}$ and $\hat{\beta} = B\hat{b}$.

Remark 1. Though D_4 appears to be unnecessary at the first glance, it is in fact essential as illustrated in the following example. Assume $Y = f((\beta_1 \otimes \alpha_1)^T X, (\beta_2 \otimes \alpha_2)^T X, (\beta_1 \otimes \alpha_2)^T X, \varepsilon)$, where $\alpha_1^T \alpha_2 = \beta_1^T \beta_2 = 0$. Then the third CTDRS direction $\beta_1 \otimes \alpha_2$ in the previous model is in D_4 and will not be able to recover if we omit D_4 in the optimization.

Remark 2. It is also necessary to point out that D_1 to D_4 is not a partition of D^k, where $D^k = \{(\alpha, \beta) \mid (\beta \otimes \alpha)^T \hat{\gamma}_i = 0, i = 1, \ldots, k\}$ is all decomposable tensors in $\mathbb{R}^{p_2} \otimes \mathbb{R}^{p_1}$ that is orthogonal to the space spanned by $\{\hat{\gamma}_1, \ldots, \hat{\gamma}_k\}$. In fact, we can easily show that

$$D_1 \cup D_2 \cup D_3 \cup D_4 \subset D^k$$

using the following example. Assume that $Y = f((\beta_1 \otimes (\alpha_0 + \alpha_1))^T X, \varepsilon)$, where $\beta_1 \in \mathcal{S}_{2,k}^{\perp}$, $\alpha_1 \in \mathcal{S}_{1,k}$, and $\alpha_0 \in \mathcal{S}_{1,k}^{\perp}$. It is easy to show that $\gamma_1 = \beta_1 \otimes (\alpha_0 + \alpha_1)$ belongs to D^1 but none of the D_1 to D_4. However, we will argue that this under-coverage is not an issue for our TDR model since it is the whole CTDRS rather than individual directions in CTDRS that we want to recover. In this example, notice that we can easily use two linearly independent tensor $\beta_1 \otimes \alpha_0$ and $\beta_1 \otimes \alpha_1$ that recovered in D_1 to D_4 to estimate γ_1.

6 Case Study

6.1 2-way Tensor Classification of TIC at High and Low Concentrations

The method was applied to a series of CSA experiments for 147 chemicals with the aim to classify these chemicals into either nontoxic or one of the 20 toxic industrial chemicals (TICs) that listed as "High Hazard TICs" on the NATO International Task Force 25. Seven replicates in each chemical category were tested. The color changes of the 147 chemicals were recorded before and 5 min after exposure at a Permissible Exposure Limits (PEL), a level may cause serious health problems after multiple exposures.

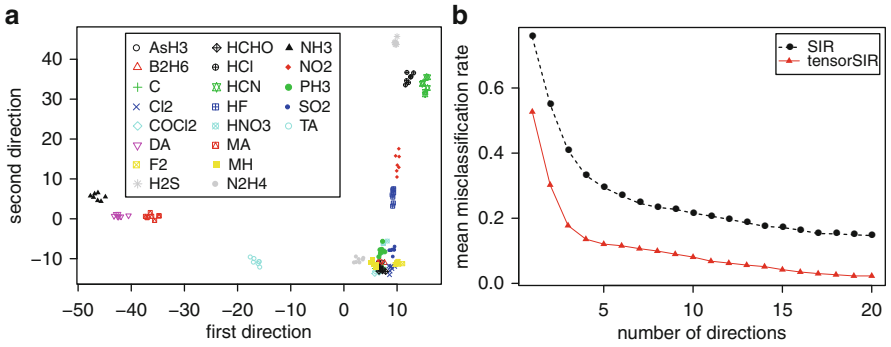


Fig. 2 Panel (a) shows the projections of 147 VCTs on the first two tensor sir directions at PEL level. Panel (b) is the average misclassification error of testing sets versus the number of tensor SIR directions.

Figure 2a is the projection of the 147 chemicals on the first and second tensor SIR directions. Though some of the chemical categories can be clearly classified using the first two tensor SIR directions, most chemical categories have no clear boundary. This clearly supports the motivation of the CSA that we need to test large amount of chemical interactions to improve the classification accuracy. To further pin down the best number of tensor SIR direction, we calculated the predicted misclassification error by randomly sample one chemical from each TIC class to form the testing set and use the rest of the chemicals to form the training set. The random partition has been done for 350 times. We plotted in Fig. 2b the average misclassification error of SIR and tensor SIR versus the number of directions. We can also see clearly that tensor SIR outperforms SIR consistently in terms of misclassification error. The minimum misclassification error obtained using SIR is 15 % while the error is reduced to 2 % by using the tensor SIR.

6.2 3-Way Tensor Classification of Bacterias

Rapid and accurate detection of pathogenic bacteria is important for determining potential medical remedies. A physician may be unable to address a bacteria infection with the appropriate antibiotic until the identity or antibiotic susceptibility of the bacteria has been determined, and consequently, sepsis remains one of the leading causes of death even among first-world nations. The detection of pathogenic bacteria is also very important for industry to contain the potential bacteria contamination. It has been shown in Carey and Suslick (2011) that the bacterial can be smelled using the CSA.

In this experiment, the color difference map of 150 pathogenic bacterials from 10 bacterial class and 14 control has been collected every 30 min using CSA with a time range from 120 to 600. For each bacterial a 3-way array is generated with

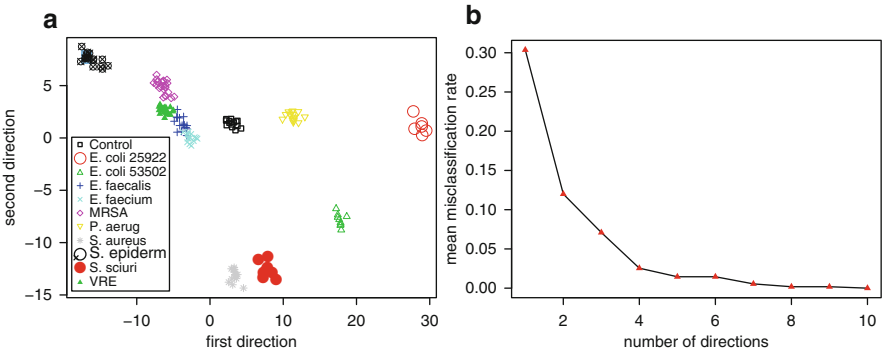


Fig. 3 Panel (a) shows the projections of the 164 different maps on the first two tensor sir directions. Panel (b) is average misclassification error of testing sets versus the number of tensor SIR directions.

the first dimension being the dye effect, second dimension being the color spectrum effect, and the third dimension being the time effect. Because the change of the color for the same analyze is not continuous as the time increase, we treat the time effect as a multivariate vector rather than a continuous variable.

Figure 3a is the projection of the 164 color difference map on the first and second tensor SIR directions. Though some of the bacterials such as E. faecalis and E. faecium cannot be clearly separated using the first two tensor SIR direction, most bacterials can be grouped well. To further evaluate the sensitivity and specificity of the CSA in bacterial detection, we separate the data by randomly sample one difference map from each class to form the testing sample and use the rest difference maps to form the training sample. Tensor SIR is applied on the training data and the misclassification error is calculated using the testing data. The procedure is repeated 50 times. In Fig. 3b, we plotted the average misclassification error of the testing set along different number of tensor SIR directions. The misclassification rate is 2.55 % with four directions and below 1 % with 7 or more directions. Figure 3b supported the biological assumption that the bacteria produce volatile organic compounds that can be well detected using the CSA. We did not compare our method with SIR for this study because the data in this study has a severe $p \gg n$ problem ($p = 1836$ and $n = 164$) that cannot be handled by SIR.

7 Conclusion

In this paper, we proposed a TDR model under the SDR framework. As all the other sufficient dimension reduction methods, our TDR model did not assume a specific relationship between the response variable and the explanatory variables. Consequently, the model can be used for a wide spectrum of scientific applications such as the one that we used in this paper.

Moreover, as discussed in Sect. 3, the TDR model can effectively reduce the parameter space by incorporating the tensor structure into the dimension reduction regression. Thus, as we demonstrated in the example in Section 6.1, it is of a great advantage for dealing with the tensor data that cannot be handled using a traditional vector analysis method.

As our model intends to keep the tensor structure of the indexes for its interpretability, we may not be able to find the exact central dimension reduction subspace. Compared to the existing method such as dimension folding method that proposed in Li et al. (2010), our method can provide a more parsimonious estimate of the central space $\mathcal{S}_{Y|X}$, and moreover, a more interpretable estimate directions.

Like all the iterative estimation approach, the SIDRA procedure proposed in this article may encounter issues typical to iterative estimation approach as discussed in Lange (2010). One major limitation of the iterative estimation approach is that the estimate may practically trapped to a local optimal estimate and fail to reach the global optimal estimate. To mitigate this issue, we advocate trying multiple starting point for iteration in each hierarchy and choose the best estimate. The proposed algorithm can be easily generated to the functional tensor data. This approach is currently under investigation and the results will be reported in the future publication.

Acknowledgment This work was supported by one National Science Foundation grant DMS 1107029 to PZ and two National Science Foundation grant DMS 1120256 and DMS 1228288 and one National Institutes of Health grant R01 DK062777 to WZ.

References

Carey, J. R. and K. S. Suslick (2011). Rapid identification of bacteria with a disposable colorimetric sensing array. *J. Am. Chem. Soc. 133*, 7571–7576.

Carey, J. R., K. S. Suslick, K. Hulkower, J. Imlay, K. Imlay, C. Ingison, J. Ponder, A. Sen, and A. Wittrig (2011). Papid identification of bacteria with a disposable colorimetric sensing array. *J. Am. Chem. Soc. 133*(19), 7571–7576.

Chen, C.-H. and K.-C. Li (1998). Can SIR be as popular as multiple linear regression? *Statist. Sin. 8*, 289–316.

Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Amer. Statist. Assoc. 91*, 983–992.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. John Wiley & Sons.

Cook, R. D. and S. Weisberg (1994). *An Introduction to Regression Graphics*. John Wiley & Sons.

Donoho, D. L. and M. Elad (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ minimization. *Proc Natl Acad Sci 100*(5), 2197–2202.

Duan, N. and K.-C. Li (1991). Slicing regression: A link-free regression method. *Ann. Statist. 19*, 505–530.

Ehmann, R., E. Boedeker, U. Friedrich, J. Sagert, J. Dippon, G. Friedel, and T. Walles (2012). Canine scent detection in the diagnosis of lung cancer: revisiting a puzzling phenomenon. *Eur Respir J 39*, 669–676.

Fan, J. and R. Li (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *Proceedings of the International Congress of Mathematicians*, pp. 595–622. European Mathematical Society.

Feng, L., C. J. Musto, J. W. Kemling, S. H. Lim, W. Zhong, and K. S. Suslick (2010). Colorimetric sensor array for determination and identification of toxic industrial chemicals. *Anal. Chem. 82*, 9433–9440.

Harshman, R. and M. Lundy (1984). The PARAFAC model for three-way factor analysis and multidimensional scaling. In *Research Methods for Multimode Data Analysis*, pp. 122–215. Praeger.

Hsing, T. and R. J. Carroll (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist. 20*(2), 1040–1061.

Kolda, T. and B. Bader (2009). Tensor decompositions and applications. *SIAM Review 51*(3), 455–500.

Lange, K. (2010). *Numerical Analysis for Statisticians (second edition)*. Springer.

Li, B., M. K. Kim, and N. Altman (2010). On dimension folding of matrix or array valued statistical objects. *Ann. Statist. 38*(8), 1094–1121.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc. 86*, 316–327.

Lim, S. H., C. J. Musto, E. Park, W. Zhong, and K. S. Suslick (2008). A colorimetric sensor array for detection and identification of sugars. *Organic Letters 10*, 4405–4408.

Merris, R. (1997). *Multilinear Algebra*. CRC Press.

Morgan, R. (2012). Thinking outside the box about screening for ovarian cancer: The nose knows! *J. Natl. Compr. Canc. Netw 10*, 795–796.

Rakow, N., A. Sen, M. Janzen, J. Ponder, and K. Suslick (2005). Molecular recognition and discrimination of amines with a colorimetric array. *Angew. Chem. Int. Ed. 44*, 4528–4532.

Rakow, N. A. and K. S. Suslick (2000). A colorimetric sensor array for odour visualization. *Nature 406*, 710–714.

Savage, N. (2012). Technology: The taste of things to come. *Nature 486*, S18–S19.

Smilde, A., R. Bro, and P. Geladi (2004). *Multi-way Analysis with Applications in the Chemical Sciences*. John Wiley & Sons.

Tucker, L. (1951). A method for synthesis of factor analysis studies. *Personnel Research Section, Report 984*(Dept. of Army).

Tucker, L. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika 31*(3), 279–311.

Zeng, P. and Y. Zhu (2010). An integral transform method for estimating the central mean and central subspaces. *J. of Multivariate Analysis 101*, 271–290.

Zhang, C., D. P. Bailey, and K. S. Suslick (2006). Colorimetric sensor arrays for the analysis of beers: a feasibility study. *J. Agric. Food Chem. 54*, 4925–4931.

Zhang, C. and K. Suslick (2005). A colorimetric sensor array for organics in water. *J. Am. Chem. Soc. 127*, 11548–11549.

Zhong, W. and K. Suslick (2012). Penalized classification for matrix predictors with application to colorimetric sensor arrays. *Technometrics*, accepted.

Zhong, W., T. Zhang, Y. Zhu, and J. Liu (2012). Correlation pursuit: forward stepwise variable selection for index model. *J. Roy. Statist. Soc. Ser. B 74*, 849–870.

Successive Standardization: Application to Case-Control Studies

Bala Rajaratnam, Sang-Yun Oh, Michael T. Tsiang, and Richard A. Olshen

Abstract In this note we illustrate the use and applicability of successive standardization (or normalization), studied earlier by some of the same authors (see Olshen and Rajaratnam, Algorithms 5(1):98–112, 2012; Olshen and Rajaratnam, Proceeding of the 1st International Conference on Data Compression, Communication and Processing (CCP 2011), June 21–24, 2011; Olshen and Rajaratnam, Annals of Statistics 38(3):1638–1664, 2010), in the context of biomedical applications. Successive standardization constitutes a type of normalization that is applied to rectangular arrays of numbers. An iteration first begins with operations on rows: first subtract the mean of each row from elements of the particular row; then row elements are divided by their respective row standard deviations. This constitutes half an iteration. These two operations are then applied successively at the level of columns, constituting the other half of the iteration. The four operations together constitute one full iteration. The process is repeated again and again and is referred to as "successive standardization." Work in Olshen and

B. Rajaratnam
Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305-4065, USA
e-mail: brajarat@stanford.edu

S.-Y. Oh
Institute for Computational & Mathematical Engineering, Stanford University, Stanford, CA 94305, USA
e-mail: sangoh@stanford.edu

M.T. Tsiang
Department of Environmental Earth System Science, Stanford University, Stanford, CA 94305, USA
e-mail: mtsiang@stanford.edu

R.A. Olshen (✉)
Department of Health Research and Policy, Stanford University School of Medicine, HRP Redwood Building, Stanford, CA 94305-5405, USA
e-mail: olshen@stanford.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_19,

Rajaratnam, Algorithms 5(1):98–112, 2012; Olshen and Rajaratnam, Proceeding of the 1st International Conference on Data Compression, Communication and Processing (CCP 2011), June 21–24, 2011; Olshen and Rajaratnam, Annals of Statistics 38(3):1638–1664, 2010 is about both theoretical and numerical properties of the successive standardization procedure, including convergence, rates of convergence, and illustrations. In this note, we consider the application of successive standardization to a specific biomedical context, that of case–control studies in cardiovascular biology. We demonstrate that successive standardization is very useful for identifying novel gene therapeutic targets. In particular, we demonstrate that successive standardization identifies genes that otherwise would have been rendered not significant in a Significance Analysis of Microarrays (SAM) study had standardization not been applied.

1 Introduction

Data in the form of large rectangular arrays of real numbers arise naturally in the biomedical sciences (such as in gene expression from protein chips), the earth and environmental sciences and many other applications. In the biomedical context, columns correspond to subjects and rows correspond to genes or gene fragments. In comparing two vectors of interest in such contexts, for instance, take two rows of such matrices that may have been measured in different scales. One can put these rows "on the same footing" by subtracting the mean of each vector followed by dividing by its standard deviation. Sometimes, the need to do a similar type of standardization occurs at the level of the entire rectangular array. For these purposes, a successive standardization procedure was proposed by colleague Bradley Efron as a possible way to put both rows and columns "on the same footing." We formally define the successive standardization procedure in the section that follows this brief introduction. The procedure was analyzed and studied by some of the authors in [1–3]. This paper continues in this line of work.

2 Successive Standardization: Theory, Methods, Illustrations

Let us now define formally the successive standardization procedure in this section. We briefly summarize its theoretical numerical properties for completeness and as a backdrop to the application that follows. To this end, we follow closely the explanation in [1–3]. Let $\mathbf{X}^{(\mathbf{0})}$ denote a rectangular array of real numbers. We first standardize this initial matrix at the level of each by row, i.e., first subtracting the row mean from each entry and thereafter dividing each entry in each row by its respective row standard deviation. The resulting matrix is denoted as $\mathbf{X}^{(\mathbf{1})}$. We defined this procedure as one iteration in the process of attempting to row and column standardize the rectangular array of real numbers or matrix. We then proceed to standardize the matrix at the level of each column, i.e., by first subtracting the

column mean from each entry (of that column) and then by dividing each entry by the respective column standard deviation. The resulting matrix is denoted as $\mathbf{X}^{(2)}$. In particular, row mean and standard deviation polishing is followed by column mean and standard deviation polishing. The process is then repeated with $\mathbf{X}^{(2)}$ and is repeated until successive standardization eventually yields a row and column standardized matrix. In particular, the successive normalizations are repeated until "convergence" which is defined as the difference in the squared Frobenius norm between two consecutive iterations being less than a pre-specified tolerance.

Following the notation in [2] closely, let $\mathbf{X}$ denote an $I \times J$ matrix with values $x \in \mathbf{R}^{IJ}$. We take coordinates $\mathbf{X}_{ij}(x) = X_{ij}$ to be iid $N(0,1)$, though this assumption can be easily relaxed. As in [1–3], we assume that $3 \leq \min(I,J) \leq \max(I,J) < \infty$.

Let $\mathbf{X} = \mathbf{X}^{(0)}$. Now

$$\bar{X}_{i\cdot}^{(0)} = \frac{1}{J}\sum_{j=1}^{J} X_{ij}; \quad \bar{X}_{\cdot j}^{(0)} = \frac{1}{I}\sum_{i=1}^{I} X_{ij}$$

$$(S_i^{(0)})^2 = \frac{1}{J}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i\cdot}^{(0)})^2 \tag{1}$$

$$= \frac{1}{J}\sum_{j=1}^{J}(X_{ij})^2 - \frac{2}{J}\sum_{j=1}^{J} X_{ij}\bar{X}_{i\cdot}^{(0)} + (\bar{X}_{i\cdot}^{(0)})^2 \tag{2}$$

$$= \frac{1}{J}\sum_{j=1}^{J}(X_{ij})^2 - 2(\bar{X}_{i\cdot}^{(0)})^2 + (\bar{X}_{i\cdot}^{(0)})^2. \tag{3}$$

$\mathbf{X}^{(1)} = [X_{ij}^{(1)}]$, where, in an obvious notation, $X_{ij}^{(1)} = (X_{ij} - \bar{X}_{i\cdot}^{(0)})/S_i^{(0)}$.

From [1–3], almost surely $(S_i^{(0)})^2 > 0$. Analogously, set $\mathbf{X}^{(2)} = [X_{ij}^{(2)}]$, where $X_{ij}^{(2)} = (X_{ij}^{(1)} - \bar{X}_{\cdot j}^{(1)})/S_j^{(1)}$, the definition of $(S_j^{(1)})^2$ being self-explanatory. As in [1–3], $(S_j^{(1)})^2 > 0$. In general, for m odd, $X_{ij}^{(m)} = (X_{ij}^{(m-1)} - \bar{X}_{i\cdot}^{(m-1)})/S_i^{(m-1)}$ with $\bar{X}_{i\cdot}^{(m-1)}$ and $(S_j^{(m-1)})^2$ defined in obvious ways. Now for m even $X_{ij}^{(m)} = (X_{ij}^{(m-1)} - \bar{X}_{\cdot j}^{(m-1)})/S_j^{(m-1)}$, with $\bar{X}_{\cdot j}^{(m-1)}$ and $(S_j^{(m-1)})^2$ defined analogously. We assume without loss that $(S_j^{(l)})^2$ and $(S_j^{(l)})^2$ is positive for all (i,j,l).

The results in [1–3] pertain both to convergence and rates thereof of the above successive standardization procedure. In particular, successive standardization is provably convergent except on a set of measure zero.

We now proceed to illustrate the pattern of convergence on matrices of different dimensions. We first consider 5-by-5 matrices. Figure 1 gives plots of the log of the squared Frobenius norms of the differences between consecutive iterates for such matrices. Four different plots have been included to demonstrate that the number of iterations required to attain convergence varies from one starting point to another. These figures describe the type of convergence patterns that are observed in the small dimensional setting from different starting values. Though different starting

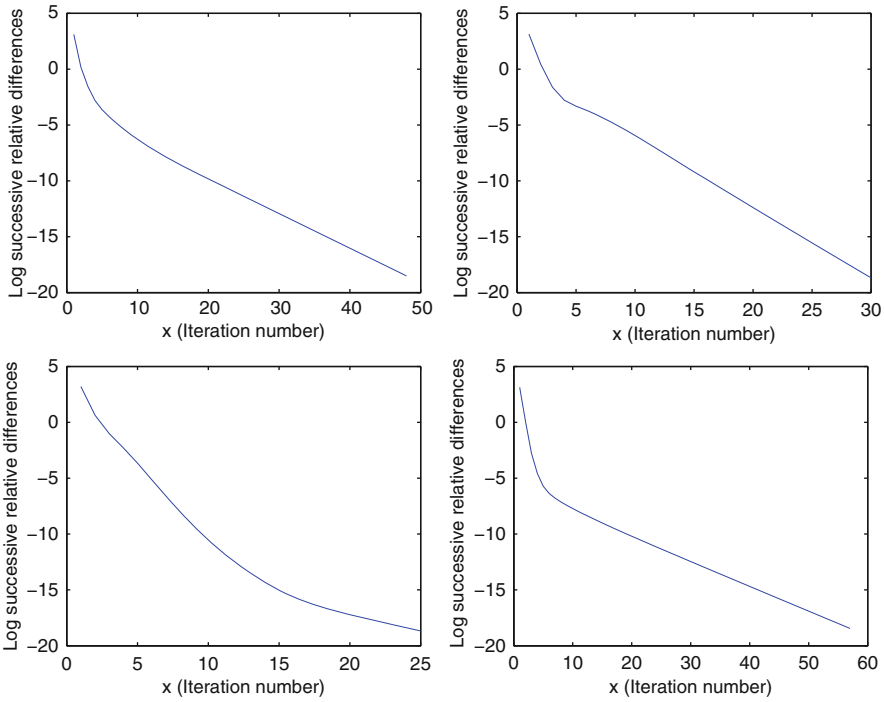


Fig. 1 Convergence patterns of successive standardization for 5-by-5 matrices. Starting values were generated from a uniform [0,1] distribution. The y-axis denotes the log of the squared Frobenius norm of the differences between consecutive iterates. The x-axis denotes iteration number.

values give slightly different convergence patterns, it is clear that the convergence is rapid. We also provide illustrations of convergence for larger (20-by-20) matrices—see Fig. 2. The convergence patterns in the 20-by-20 case are more homogeneous. Once more it is clear that regardless of dimension of the matrix, convergence of successive standardization is rapid. This rapidity of convergence to the set of fixed points of successive standardization can be visualized when $I = J = 3$, see Fig. 3.

The work in [1] also concerns a related question, which is whether the convergence phenomenon observed with successive normalization will still occur if simultaneous normalization is undertaken instead. More specifically, simultaneous normalization entails row and column mean polishing and row and column standard deviation polishing all done at once:

$$X_{ij}^{(t+1)} = \frac{X_{ij}^{(t)} - \bar{X}_{i.}^{(t)} - \bar{X}_{.j}^{(t)}}{S_{i.}^{(t)} S_{.j}^{(t)}}.$$

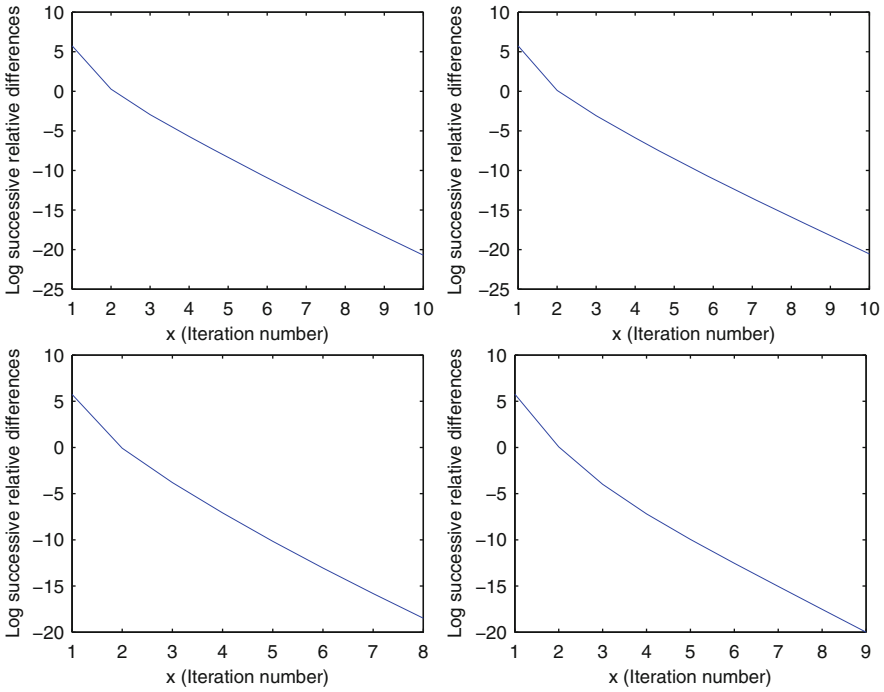


Fig. 2 Convergence patterns of successive standardization for 20-by-20 matrices. Starting values were generated from a uniform [0,1] distribution. The y-axis denotes the log of the squared Frobenius norm of the differences between consecutive iterates. The x-axis denotes iteration number.

It can be shown [1] through counter example that the simultaneous normalization algorithm does not converge. We provide a few plots to illustrate this lack of convergence (see Fig. 4).

These counterexamples do not, however, shed light on the possibility of convergence for simultaneous normalization. The numerical work undertaken in this regard indicates that convergence was not seen for any of the examples that were tried.

3 Application of Successive Standardization in Cardiovascular Biology

We now proceed to investigate the merits of using successive standardization in the context of cardiovascular biology and cardiovascular disease. We focus first on a case–control study, where the specific disease under investigation is ischemic cardiomyopathy (ICM). We shall then consider data on dilated cardiomyopathy (DCM).

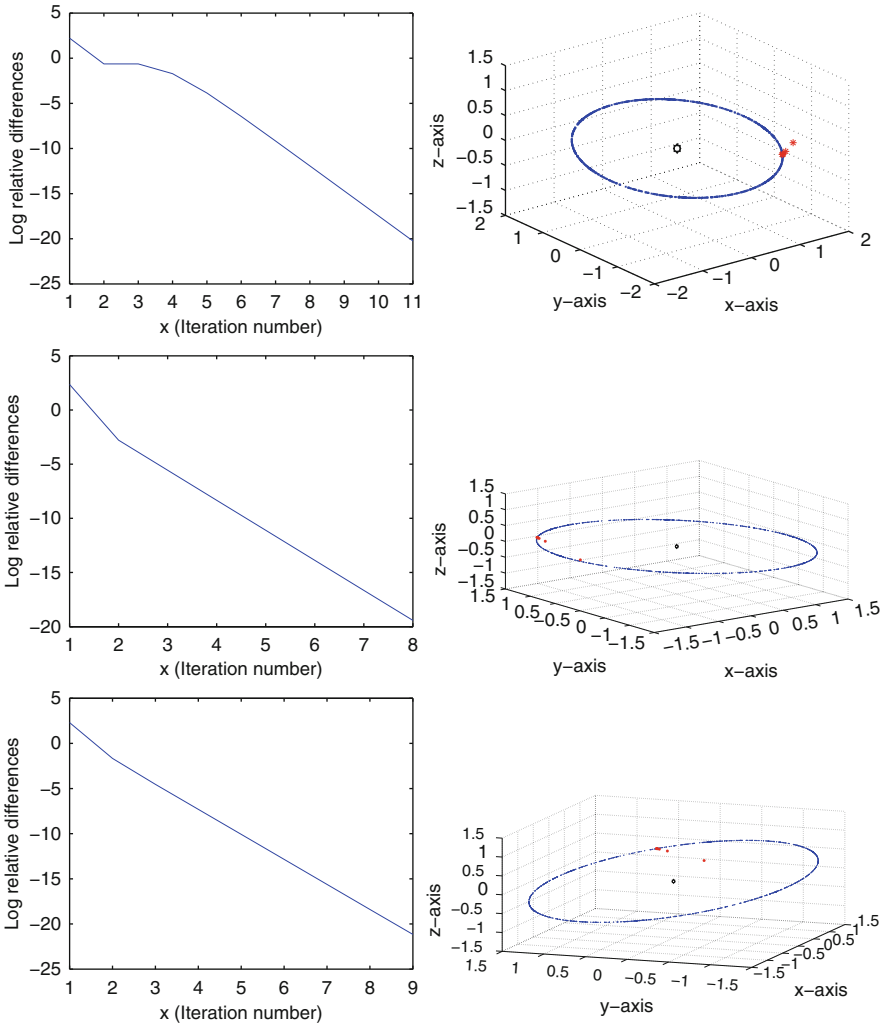


Fig. 3 Illustration of convergence to set of fixed points in the $I = J = 3$ case. Note that due to constraints implied by the mean and standard deviation for each row and column only 3 distinct entries appear in the limit of the sequence of iterates. These 3 distinct values appear in each column of the limit and are therefore sufficient to illustrate each limit. The x, y, z axes in the right panel denote the coordinates of the 3rd column of each iterate. The black circle indicates the origin as given by (0,0,0). The blue circle indicates the set of fixed points in the $I = J = 3$ case. The red points indicate the 3rd column of successive iterates with initial values generated from a uniform [0,1] distribution.

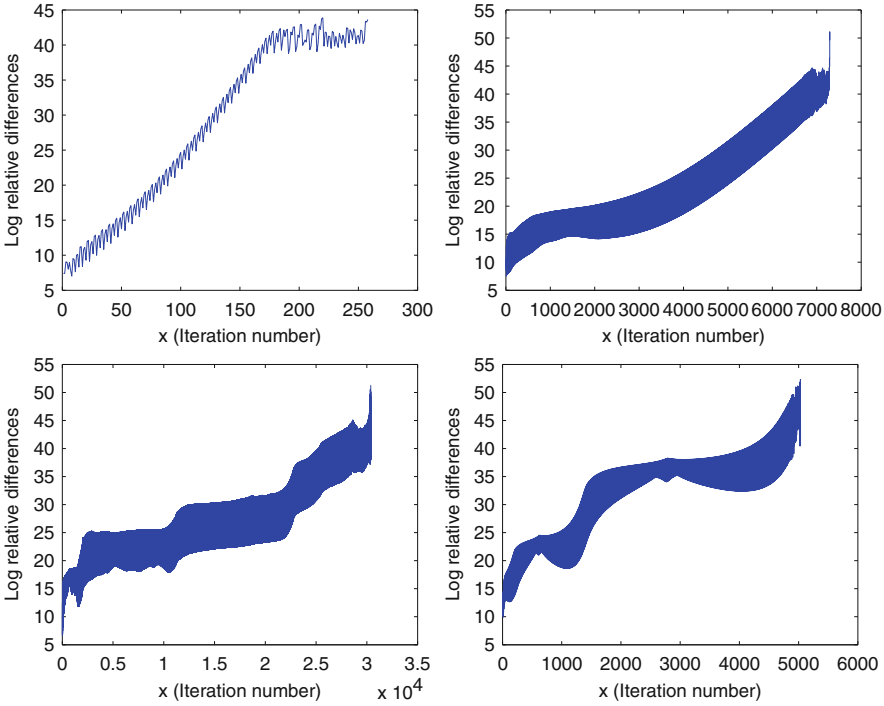


Fig. 4 Illustration of simultaneous normalization and divergence of procedure. Starting values were generated from a uniform [0,1] distribution. The y-axis denotes the log of the squared Frobenius norm of the differences between consecutive iterates. The x-axis denotes iteration number.

3.1 Ischemic Cardiomyopathy

We consider dataset GSE5406 from the Gene Expression Omnibus. This dataset was gathered from 16 normal individuals and 108 with ischemic cardiomyopathy. The number of genes in the study is 22283. In order to assess if successive normalization reveals any new genes, we use the Significance Analysis of Microarrays (SAM) method [4] to see if with and without standardization yields different results. We shall also refer to the dataset prior to standardization as the "raw" data and the standardized data as its "normalized" version. The number of iterations required to obtain the standardized data from the raw data is 12 for the normal group and 10 iterations each for the ICM and DCM groups. The starting value is fixed in this example as given by the original dataset. In order to run SAM the number of genes was reduced by approximately half to 11,820. This reduction was for both illustration and identifiability purposes, and also serves to make the SAM analysis computationally tractable. Recall that probe sets on the HG-U133A platform measure abundance of different (sets of) sequences. The majority of probe sets are

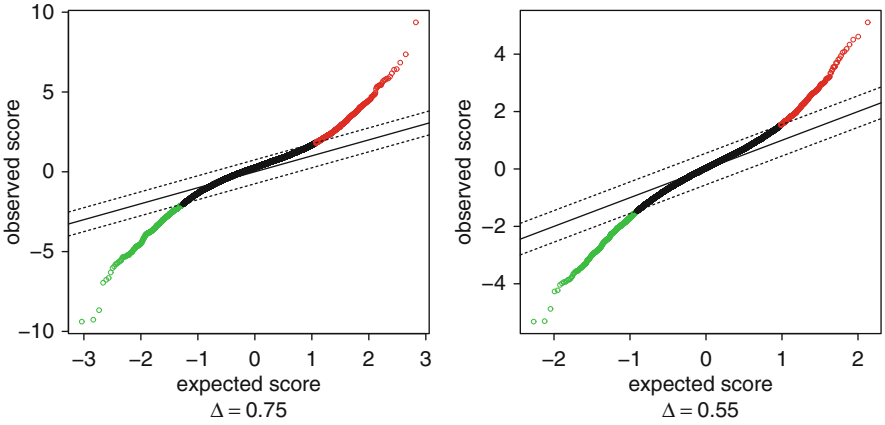


Fig. 5 SAM Q-Q plots for ICM analysis before (left) and after standardization (right). The red points denote the genes that are significantly overexpressed, and the green points denote those that are significantly underexpressed. Note that the median FDR is 0.05543605 for the raw data (delta = 0.75) and a median FDR of 0.05970533 for the normalized data (delta = 0.55).

Table 1 Summary of results from the SAM method for the data on ischemic cardiomyopathy.

	Raw	Standardized
Number of significant genes recorded	1507	1233
Number of genes significantly overexpressed	897	499
Number of overexpressed genes not identified in alternative method	403	5
Number of genes significantly underexpressed	610	734
Number of underexpressed genes not identified in alternative method	56	180

designed to measure expression levels of particular genes; however, some probe sets measure abundance of expression levels of a family of genes that share a common subsequence or a set of common subsequences.[1] For our analysis we have retained probe sets that are designed to measure expression levels of particular genes only (these are probe sets having "_at" designations). After this filtering, 11820 genes remain. The samr() function in library(samr) in R was implemented on both the raw and normalized data with the number of permutations set at 1,000.

The SAM analysis led to constructing a delta-table using samr.compute.delta.table() for values of delta (a SAM parameter) between 0 and 2, incrementing by 0.05. In order to compare the lists of genes that are called significant between the two datasets, we chose a median false discovery rate (FDR) of 0.05543605 for the raw data (delta =0.75) and a median FDR of 0.05970533 for the normalized data (delta

[1] http://media.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf

Table 2 Summary of results from the SAM method for the data on dilated cardiomyopathy.

	Raw	Standardized
Number of significant genes recorded	1985	1337
Number of genes significantly overexpressed	744	627
Number of overexpressed genes not identified in alternative method	188	71
Number of genes significantly underexpressed	1241	710
Number of underexpressed genes not identified in alternative method	532	1

= 0.55). The SAM Q-Q plots when we control the median FDR at their respective levels are shown in Fig. 5. The results are recorded in Table 1:

Table 1 indicates that of the 499 significantly over expressed genes after standardization, five genes were not considered significant in the analysis before normalization. The gene symbols for these five are: PIK3R4, ODC1, CAMKK2, INSL5, PMP2. We note that the gene ODC1 (ornithine decarboxylase), for instance, has featured in a study where its expression level was up regulated. The study concerned ischemia-reperfusion (I/R) injury in an animal (rat) model (see http://www.ncbi.nlm.nih.gov/pubmed/19100015?dopt=Abstract).

3.2 Dilated Cardiomyopathy

The analysis for dilated cardiomyopathy is similar to the one above for ischemic cardiomyopathy. We give the specific details of the analysis for the sake of completeness. We once more consider dataset GSE5406 from the Gene Expression Omnibus. This dataset contains data from 16 normal individuals and 86 from those with dilated cardiomyopathy (DCM). The number of genes in the study is the same, i.e., 22,283. In order to assess if successive normalization reveals any new genes, we once again use the Significance Analysis of Microarrays (SAM) method to see if with and without standardization yield different results. In order to run SAM the number of genes was once more reduced by approximately half to 11820. As mentioned in the analysis of ischemic cardiomyopathy, this reduction is for illustrative purposes only and renders the SAM analysis more tractable computationally. The samr() function in library(samr) in R was implemented on both the raw and normalized data with the number of permutations set at 1,000.

The SAM analysis led to constructing a delta-table using samr.compute.delta.table() for values of delta (a SAM parameter) between 0 and 2, incrementing by 0.05. In order to compare the lists of genes which are called significant between the two datasets, we chose a median FDR of 0.0569515 for the raw data (delta = 0.8) and a median FDR of 0.05233577 for the normalized data (delta = 0.5). The SAM Q-Q plots when we control the median FDR at their respective levels are shown in Fig. 6. The results are recorded in Table 2:

Table 2 indicates that of the 627 significantly over expressed genes after standardization, 71 genes were not considered significant in the analysis before

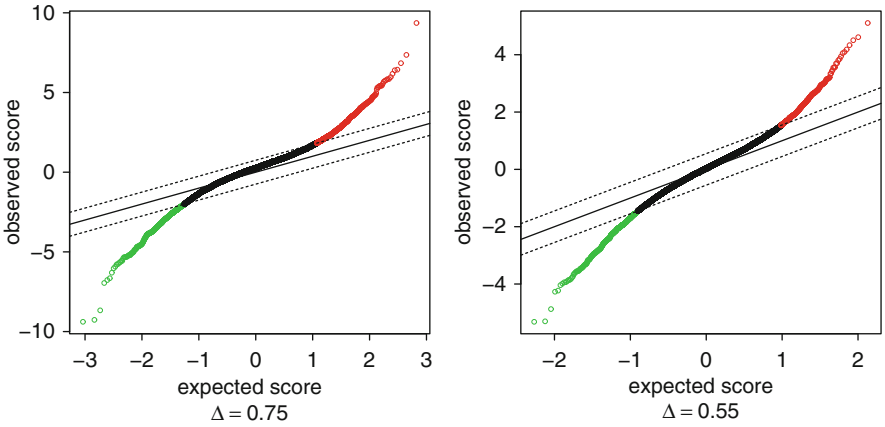


Fig. 6 SAM Q-Q plots for DCM analysis before (left) and after standardization (right). The red points denote the genes that are significantly overexpressed, and the green points denote those that are significantly underexpressed. Note that the median FDR is 0.0569515 for the raw data (delta = 0.8) and a median FDR of 0.05233577 for the normalized data (delta = 0.5).

normalization. Moreover, of the 710 significantly under expressed genes after standardization, one particular gene was not considered significant in the analysis before normalization. The gene symbol for this particular gene is GZMB (granzyme B). We note that the gene GZMB (granzyme B) has been connected to dilated cariomyopathy in the literature (http://www.ncbi.nlm.nih.gov/pubmed/10398167?dopt=Abstract).

4 Concluding Remarks

The goal of this note is to demonstrate that successive standardization has not only theoretical properties but also potentially important practical applications, especially in the context of genomics. The comparison of significantly expressed genes pre- and post-standardization in a case–control setting reveals that standardization can lead to the identification of novel genes that would otherwise have been omitted. This does not mean to say that standardized data should replace the original data, but rather that together they are complementary and can lead jointly to novel therapeutic targets. This is affirmed by the finding that standardized data can fail to discover genes that raw data is able to and vice versa. Future work in this regard will be in the direction of applying standardization to other case–control studies outside the context of cardiovascular biology. This includes applying the technique to datasets on breast and other types of cancer.

Acknowledgments The authors thank colleague Bradley Efron for introducing them to the original problem. Bala Rajaratnam was supported in part by National Science Foundation grants DMS-CMG 1025465, AGS-1003823, DMS-1106642 and grants SUWIEVP10-SUFSC10-SMSCVISG0906. Richard Olshen was supported in part by grants 4R37EB002784-35 (an NIH MERIT award), 1U19AI090019-01 and UL1RR025744. Sang Oh was supported in part by NSF-DMS-CMG 1025465.

References

[1] Olshen, R.A. and Rajaratnam, B. (2012), Successive normalization of rectangular arrays, *Algorithms*, 5(1), 98–112. doi:10-3390/a5010098

[2] Olshen, R.A. and Rajaratnam, B. (2011), Successive normalization of rectangular arrays, *Proceeding of the 1st International Conference on Data Compression, Communication and Processing (CCP 2011)*, Palinuro, June 21–24, 2011. doi:10.1109/CCP2011.48

[3] Olshen, A. and Rajaratnam, B. (2010), Successive normalization of rectangular arrays, *Annals of Statistics* 38, No. 3, 1638–1664. doi:10.1214/09-AOS743

[4] Tusher, V. Tibshirani, R. and Chu, G. (2001), Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc. Natl. Acad. Sci. USA.*, 98:5116–5121.

Part V
Survival Analysis

Quantification of PFS Effect for Accelerated Approval of Oncology Drugs

Cong Chen and Linda Z. Sun

Abstract By the accelerated approval (AA) mechanism (Code of Federal Regulations- 21 CFR 314 and 601. Accelerated Approval Rule, 1992), the FDA may grant approval of drugs or biologic products that are intended to treat serious or life-threatening diseases using a surrogate endpoint that is *reasonably likely* to predict clinical benefit. In oncology, progression-free-survival (PFS) is increasingly used as such a surrogate of overall survival (OS) in Phase III confirmatory trials. Improved understanding on how to deal with the PFS endpoint in trial conduct and data analysis has mitigated some regulatory concerns about this endpoint. However, a glaring gap still exists as how to determine whether the outcome from a registration trial with PFS as the primary endpoint at the time of analysis is *reasonably likely* to predict a clinical benefit as normally reflected through an effect on OS. Since there is no guidance on this, regulatory agencies tend to look for a compelling PFS effect coupled with an OS effect in the right direction without specification of the effect sizes and significance levels. To address this issue, we propose a synthesized approach that combines the observed OS effect and the estimated OS effect from the PFS data to explicitly test the implicit OS hypothesis at the time of primary analysis. The proposed approach is applied to hypothetical Phase III trials in metastatic colorectal cancer and adjuvant colon cancer settings using the relationships between OS effect size and PFS effect size established from historical data. Prior information on such a historical relationship is frequently cited by relevant decision makers during regulatory reviews for drug approval. However,

C. Chen (✉) • L.Z. Sun
Biostatistics and Research Decision Sciences, Merck Research Laboratories (MRL), UG1C-46, 351 N. Sumneytown Pike, Upper Gwynedd, PA 19454, USA
e-mail: cong_chen@merck.com; linda_sun@merck.com

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_20,

the information is rarely fully accounted for in the actual (mostly qualitative) decision-making process. Our approach provides a simple analytic tool for deriving a more quantitative decision. It is clear that the design based on our approach may have a larger sample size than a conventional trial with PFS as the primary endpoint, but directly address the elusive OS question that a conventional PFS trial cannot, no matter how good a surrogate endpoint PFS is.

1 Introduction

The conventional endpoint for a Phase III confirmatory trial in oncology is overall survival (OS), defined as time from randomization to death due to any cause. In recent years, there is an increasing interest in progression-free-survival (PFS), i.e., time from randomization to disease progression (or recurrence of disease in adjuvant setting in which case the endpoint of same interest would be disease-free-survival) or death due to any cause, whichever comes first. There are a few major drivers behind it. First, the new generation of oncology drugs under development tends to have a better safety profile than traditional chemotherapies, which makes accelerated approval based on PFS more justifiable, as the new generation oncology drugs are unlikely to be so toxic that patients' survival may be harmed by the side effects of a drug. Second, there are inevitable situations where cross-over (or cross-in) after disease progression confounds the analysis of OS, whereas PFS arguably captures the true interventional effect of an experimental treatment on tumor burden. Third, some cancer diseases have a long natural history, making it less feasible or ethical to conduct time-consuming survival trials. Fourth, disease progression tends to correlate with patient's quality of life, and in some situations may arguably be considered an indirect or even direct measure of clinical benefit especially when clinical judgment (in a blinded fashion) is incorporated into the definition, although a subject-level correlation doesn't necessarily translate into a correlation in treatment effects. Fifth, as we better understand the issues involving PFS, the conduct and analysis of clinical trials with PFS as an endpoint should improve. Last but not least, patients participating in trials are limited and societal investment is under constraint. Judicious use of PFS helps make drug development more efficient so that unmet medical needs can be met sooner rather than later although, in absence of real breakthroughs in cancer treatment, the meaningfulness of small improvements in PFS (and for this matter OS) is always questionable.

There has been a lot of statistical research investigating whether OS can be replaced with PFS, i.e., whether PFS is a surrogate endpoint for OS in the statistical sense. For example, Burzykowski et al. [2] showed that treatment effect on disease-free-survival (after 3 years of median follow-up) reliably predicts treatment effect on overall survival (after 5 years of median follow-up) for adjuvant colorectal cancer studies, supporting a previous finding by Sargent et al. [3]. Similarly, Tang et al. [4] showed that PFS effect and OS effect have a reliable relationship in metastatic colorectal cancer and Miksad et al. [5] showed that PFS effect and OS effect have

a reliable relationship in advanced breast cancer. However, can PFS be considered a surrogate endpoint of OS for future clinical trials in the same disease settings? It remains controversial from a clinical standpoint [6] (also see [7] for discussions in different disease settings). From a statistical standpoint, a surrogacy measure is often continuous in nature. An endpoint is considered a surrogate if such a measure (or measures) meet certain threshold(s). However, determination of the threshold(s) is judgmental and lacks consensus. In this sense, it is not about whether an endpoint is a surrogate or not; it is more about to what extent (or degree) an endpoint is a surrogate. Discussions about various surrogacy measures can be found in [8, 9]. A less ambitious application of one such measure (proportion of treatment effect explained) for hypothesis generation, as opposed to surrogacy validation, can be found in [10].

In this paper, we address the endpoint issue from a slightly different angle in light of the FDA requirement for accelerated approval [1], which is safeguarded with a Phase IV commitment of directly demonstrating clinical benefit, and regulatory rationales from both the EMA and FDA on recent oncology drug approvals (available in the public domain). It is apparent therein (also see [11]) that the definition of a "surrogate" endpoint is focused on whether the treatment effect on the PFS endpoint is *reasonably likely* to predict a clinical benefit as normally reflected through an effect on OS, which is a daunting task given various issues associated with this endpoint (a comprehensive discussion can be found in [12]). Statistical validation of PFS as a surrogate endpoint for OS (by some measure yet to be determined with consensus) certainly helps, but is not mandatory in this context, since it does not directly address the question: how clinically and statistically significant should a PFS effect be to meet the regulatory requirement? It relies on the strength of evidence in the ongoing trial and the historical relationship between PFS effect and OS effect, and this problem is less investigated in the statistical literature. The closest approach may be found in [13, 14] where surrogate threshold effect (STE), or minimum treatment effect on a surrogate marker required to predict a nonzero positive treatment effect on OS in a future trial, was proposed. The establishment of STE for PFS helps with a regulatory decision but does not directly address the regulatory issue. Besides, it is unclear how to incorporate OS effect into the metrics. In the absence of a solid statistical argument, regulatory agencies tend to look for a compelling PFS effect and an OS effect in the right direction without specification of the effect sizes and significance levels. As a result, regulatory decisions are subject to judgmental bias. This paper attempts to fill the gap by proposing a more objective approach.

The historical relationship between PFS effect size and OS effect size in the same disease setting holds the key to successful estimation of the OS effect based on PFS data. Prior information on the historical relationship is frequently cited by relevant decision makers during regulatory reviews for drug approval. However, the information is rarely fully accounted for in the actual (mostly qualitative) decision-making process. In this paper, we will quantify the impact of the historical relationship on the estimation of OS effect from PFS data. We will also show how the OS data from the same trial can be further incorporated into a joint test-statistic

for testing OS effect in a synthesized approach. As a result, the heuristic regulatory requirement on both PFS effect and OS effect becomes a testable hypothesis; significant clinical benefit may be claimed, in light of current and historical data, if the one-sided p-value from the joint test-statistic is less than 2.5% (say). Statistical incorporation of historical data into the drug approval process is nothing new in the regulatory environment. It is reflected in the development of medical devices where a Bayesian framework is often applied (see [15] for FDA draft guidance). Design and analysis of a non-inferiority trial also involves the estimation of the historical effect of the active comparator (see [16] for ICH guidance). Our proposed approach follows the regulatory precedent set in these fields in the sense of utilizing historical data to make a quantitative decision. Although our work is motivated by oncology, the idea in the proposed approach is equally applicable to other therapeutic areas.

The rest of the paper is organized as follows: Sect. 2 provides the joint test-statistic for OS effect and illustrates its characteristics in a hypothetical example. Section 3 provides sample size calculations for the study design of a clinical trial using this joint test-statistic. Because the main purpose of this paper is to advocate a higher-level concept, we try to keep the technical details to minimum in the main presentation. Extensions and discussion are provided in Sect. 4.

2 Estimation of OS Effect

2.1 Set-up

Consider a hypothetical Phase III oncology trial with 1:1 randomization that uses PFS as the primary endpoint for possible accelerated approval of a new treatment. Suppose that the point estimate of hazard ratio (HR; treatment vs. control) is 0.6 in PFS based on 400 events (target number of PFS events), and is 0.75 in OS based on 100 events (immature) at the time of primary analysis. A definite demonstration of unequivocal clinical benefit may be delayed until the OS data from this trial is mature, or in a separate trial as part of the Phase IV commitment. The corresponding one-sided p-value would be <0.0001 for the PFS data and 0.075 for the OS data based on the regular Z-test statistics. Is this outcome positive enough that an accelerated approval may be granted? To properly answer this question, we need a reliable estimate of the relationship between OS effect size and PFS effect size in the same disease setting from historical data. We assume that the relationship can be approximated by a simple linear model on log(HR) scale. For simplicity, we also assume that the intercept term in the linear model is zero. This is equivalent to the assumption that if the treatment effect is nonexistent in one endpoint it will not show up in the other endpoint. Define $\hat{\gamma}$ as the mean estimate of the ratio of OS effect to PFS effect in the log hazard ratio scale, and $\hat{\sigma}_\gamma$ (>0) as the corresponding estimate of the standard error. The historical relationship in effect size is adequately captured in the doublet $(\hat{\gamma}, \hat{\sigma}_\gamma)$ under the asymptotic normality assumption. Further, let $\hat{\Delta}_{OS}$

and $\hat{\Delta}_{PFS}$ be the respective log(HR) estimates of OS and PFS effect from the current Phase III trial of interest. Their respective variance estimates are $\hat{\sigma}^2_{OS}$ and $\hat{\sigma}^2_{PFS}$, and an estimate of the correlation between $\hat{\Delta}_{OS}$ and $\hat{\Delta}_{PFS}$ is ρ which is readily available from an appropriate resampling analysis or from the SAS procedure PROC PHREG for multiple failure data analysis by invoking, e.g., the WLW option [17].

It is not the focus of this paper on the estimation of $(\hat{\gamma}, \hat{\sigma}_\gamma)$. However, it can never be over-emphasized that the historical studies should be examined for the constancy (and range of variability) of effect size ratio, which depends on the timing of analyses for PFS and OS, among other considerations [12]. Good estimates come from appropriate meta-analysis (see [18, 19] for general reference on meta-analyses), with the assistance of sensitivity analyses. Instead, we will use existing estimates in the literature for illustration of the proposed approach. In the metastatic colorectal cancer disease setting [4], the effect size ratio is estimated as the regression slope of OS effect on PFS effect for several randomized clinical trials with moderate to large size. It was estimated to be 0.54 with standard error of 0.10 on the hazard reduction scale. The estimate of intercept was negligible, and inclusion of it or not in the linear model did not change the point estimate of the regression slope. Based on these findings, the authors concluded that a 10% hazard reduction in PFS for a novel therapy would yield an estimated $5.4\% \pm 1\%$ hazard reduction in OS. No estimates were provided on the log(HR) scale. However, because hazard reduction is a good approximation of the log(HR), We will use (0.54, 0.10) as a starting point for $(\hat{\gamma}, \hat{\sigma}_\gamma)$ for illustration purpose in this disease setting. Of note, based on a point estimate of 0.54, the estimated OS effect from the PFS effect in the hypothetical trail described at the beginning of this section is HR = 0.76, which is close to the observed OS effect (HR = 0.75) (i.e., the relationship in effect size is consistent with what is established from historical data). In an adjuvant colon cancer setting, Sargent et al. [3] estimated the regression slope to be 0.89 with standard error of 0.061 on hazard ratio scale for trials of fluorouracil-based regimes. The intercept was estimated to be 0.12, which implies $\text{HR(OS)} = 0.12 + 0.89\text{HR(PFS)}$ or $1 - \text{HR(OS)} = -0.01 + 0.89(1 - \text{HR(PFS)})$, i.e., the intercept term would be negligible at -0.01 on hazard reduction scale under the same regression slope. This is consistent with findings from [2] in that the regression slope was estimated to be 0.90 on the log(HR) scale for trials of different regimens whereas the incept term is negligible at 0.03. Based on these data, the clinical community generally believes that PFS can be used to replace OS for future trials in adjuvant colon cancer. We will use (0.89, 0.061) as a starting point for $(\hat{\gamma}, \hat{\sigma}_\gamma)$ for illustration purpose in this disease setting. Of note, the observed OS effect from the hypothetical trial would be smaller than that estimated from the PFS effect if $\hat{\gamma}$ were indeed equal to 0.89. Once a reliable estimate of $(\hat{\gamma}, \hat{\sigma}_\gamma)$ is available, how should we evaluate the strength of evidence on the OS endpoint with the observed PFS effect and OS effect from the hypothetical trial?

2.2 Estimators of OS Effect

We consider two estimators of the OS effect, one based on $\hat{\Delta}_{PFS}$ alone and the other based on both $\hat{\Delta}_{PFS}$ and $\hat{\Delta}_{OS}$. By the definition of $\hat{\gamma}$, a natural estimate of the OS effect from the historical relationship and $\hat{\Delta}_{PFS}$ is

$$\hat{\Delta}_P = \hat{\gamma}\hat{\Delta}_{PFS} \tag{1}$$

Based on (1), the same inference as in [4] on OS effect for a given $\hat{\Delta}_{PFS}$ can be easily made. The overall variance of $\hat{\Delta}_P$ is

$$Var(\hat{\Delta}_P) = \hat{\gamma}^2\hat{\sigma}^2_{PFS} + \hat{\Delta}^2_{PFS}\hat{\sigma}^2_{\gamma} + \hat{\sigma}^2_{\gamma}\hat{\sigma}^2_{PFS} \tag{2}$$

The distribution of $\hat{\Delta}_P$, as a product of two random variables, is not normal in general but behaves like a normal distribution especially when $\hat{\gamma}$ has a nonzero mean and a very small variance. A Z-test statistic (denoted to be $\hat{Z}_P$) can be naturally formed from $\hat{\Delta}_P$ and $Var(\hat{\Delta}_P)$ under the normal distribution assumption. The estimated OS effect based on PFS data alone is considered statistically significant at the 2.5% (one-sided) level if

$$\hat{Z}_P = \hat{\Delta}_P / \sqrt{Var(\hat{\Delta}_P)} < Z_{0.025} \tag{3}$$

where $Z_{(.)}$ denotes the quantile of a standard normal distribution at the corresponding level. Notice that the significance level is fixed at the 2.5% (one-sided) level for ease of presentation. The appropriate level that might be considered acceptable for drug approval should depend on whether this is an isolated result or one of two or more trials providing similar results, and various other considerations. Observe that $Var(\hat{\Delta}_P) > \hat{\gamma}^2\hat{\sigma}^2_{PFS}$, which implies that (except for the trivial case of $\hat{\sigma}_{\gamma} = 0$) numerically

$$\hat{\Delta}^2_P / \mathrm{Var}(\hat{\Delta}_P) = (\hat{\gamma}\hat{\Delta}_{PFS})^2 / \mathrm{Var}(\hat{\Delta}_P) < (\hat{\gamma}\hat{\Delta}_{PFS})^2 / (\hat{\gamma}^2\hat{\sigma}^2_{PFS}) = \hat{\Delta}^2_{PFS}/\hat{\sigma}^2_{PFS}$$

regardless of $\hat{\gamma}$. This shows that, under the normal distribution assumption for $\hat{Z}_P$, the estimated OS effect from PFS will always be less statistically significant than the PFS effect. This observation is consistent with intuition that, due to uncertainty in the relationship between OS effect and PFS effect, the PFS endpoint needs to cross a higher bar (i.e., p-value <2.5%) for the estimated OS effect to be statistically significant at the 2.5% (one-sided) level. Similarly, $\hat{\Delta}^2_P / Var(\hat{\Delta}_P) < \hat{\gamma}^2/\hat{\sigma}^2_{\gamma}$ and $\hat{\Delta}_P$ will be less statistically significant than $\hat{\gamma}$. This implies that, again, consistent with intuition, OS effect cannot be reliably estimated from PFS effect in absence of a solid historical relationship between the two.

With $\hat{\Delta}_{OS}$ available, which is often the case in any Phase III oncology trial, a weighted estimator of the OS effect that naturally incorporates $\hat{\Delta}_{OS}$ is

$$\hat{\Delta}_J = w\hat{\Delta}_{OS} + (1-w)\hat{\Delta}_P. \tag{4}$$

As mentioned previously, the estimated correlation between $\hat{\Delta}_{OS}$ and $\hat{\Delta}_{PFS}$ is ρ, hence the covariance of $\hat{\Delta}_{OS}$ and $\hat{\Delta}_p$ is $\hat{\gamma}\rho\hat{\sigma}_{OS}\hat{\sigma}_{PFS}$. The variance of $\hat{\Delta}_J$ is estimated to be

$$\begin{aligned} Var(\hat{\Delta}_J) &= w^2 Var(\hat{\Delta}_{OS}) + (1-w)^2 Var(\hat{\Delta}_P) + 2w(1-w)Cov(\hat{\Delta}_{OS}, \hat{\Delta}_p) \\ &= w^2\hat{\sigma}_{OS}^2 + (1-w)^2 Var(\hat{\Delta}_P) + 2w(1-w)\hat{\gamma}\rho\hat{\sigma}_{OS}\hat{\sigma}_{PFS} \end{aligned} \tag{5}$$

An inverse-variance weighted estimate of w is

$$w = \mathrm{Var}(\hat{\Delta}_P)/(\mathrm{Var}(\hat{\Delta}_P) + \hat{\sigma}_{OS}^2). \tag{6}$$

A more complicated joint estimate of the OS effect may be obtained by maximum likelihood estimation (MLE) [20]. It is not our focus to find an optimal estimate of the OS effect and its statistical properties in this paper. We will use the one based on w from (6), which is easy to understand and robust to model misspecifications. Again, $\hat{\Delta}_J$ and $Var(\hat{\Delta}_J)$ naturally form a joint Z-test statistic (denoted to be $\hat{Z}_J$) under asymptotic normality assumptions. The estimated OS effect is considered statistically significant at the 2.5% (one-sided) level if

$$\hat{Z}_J = \hat{\Delta}_J / \sqrt{Var(\hat{\Delta}_J)} < Z_{0.025} \tag{7}$$

We apply $\hat{Z}_P$ and $\hat{Z}_J$ to the analysis of the hypothetical trial in the next section. The correlation (ρ) between $\hat{\Delta}_{PFS}$ and $\hat{\Delta}_{OS}$ is assumed to be 0.5 for all scenarios. Of note, purely from a statistical standpoint, a negative trend on the treatment effect in either PFS or OS does not automatically rule out a positive finding based on the joint test-statistic. However, such a contradictory outcome, should it ever occur, will inevitably call into question the applicability of $(\hat{\gamma}, \hat{\sigma}_\gamma)$ and the integrity of the trial conduct. Whenever possible, the assumption on historical relationship needs to be checked from trial data before a conclusion can be made.

2.3 Illustration

Depending on the similarity in disease settings, comparators used in the control arms, toxicity profiles and biological mechanisms of action for study drugs and standard of care, and various other considerations, historical data may need to be discounted or even discarded before it is incorporated into design and analysis of a

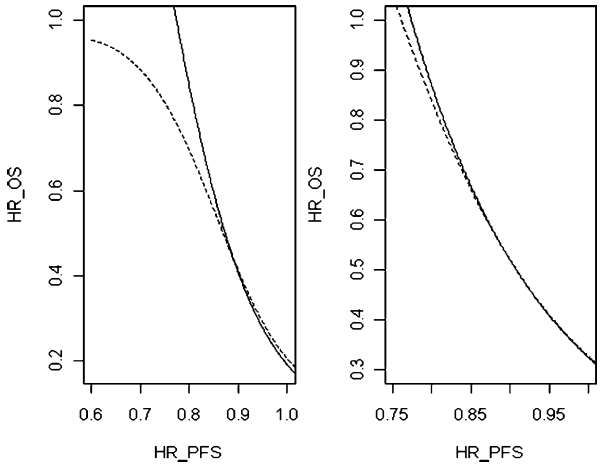


Fig. 1 Boundaries of rejection regions on hazard ratio scale based on the joint test-statistic ($\hat{Z}_J$ of 7). The solid line refers to $r = 1$ (no inflation) and the dashed line refers to $r = 2$ (doubling). The left panel is for metastatic colorectal cancer with $(\hat{\gamma}, \hat{\sigma}_\gamma) = (0.54, 0.1)$ and the right panel is for adjuvant colon cancer with $(\hat{\gamma}, \hat{\sigma}_\gamma) = (0.89, 0.061)$

future trial. One way to discount $(\hat{\gamma}, \hat{\sigma}_\gamma)$ is to inflate $\hat{\sigma}_\gamma$ by a factor of r (>1), the upper limit of which may be set at 2 (i.e., doubling of $\hat{\sigma}_\gamma$). Another way to discount $(\hat{\gamma}, \hat{\sigma}_\gamma)$ is to reduce $\hat{\gamma}$ by a certain factor so that a greater PFS effect would be needed to predict the same OS effect. For illustration purpose, we will only consider inflation of $\hat{\sigma}_\gamma$.

What kind of OS effect and PFS effect from the hypothetical trial would consist of a positive finding based on the joint test-statistic at the 2.5% (one-sided) level? Figure 1 provides boundaries of rejection regions for the empirical treatment effects on hazard ratio scale (i.e., $\exp(\hat{\Delta}_{OS})$ and $\exp(\hat{\Delta}_{PFS})$ in the two colorectal disease settings. The boundaries are confined to hazard ratio <1 for both endpoints to exclude a negative trend. As expected, the configuration generally depends on the standard error of the regression slope (via the inflation factor r). For example, when the HR for OS is 0.8 under the metastatic colorectal cancer setting, the HR for PFS needs to be 0.8 at $r = 1$ or 0.75 at $r = 2$ to have a positive outcome. However, if the OS effect has a strong positive trend (e.g., <0.6) in the metastatic colorectal setting (left panel) or <0.8 in the adjuvant colon cancer setting (right panel), the impact of r would be minimal as the boundaries for $r = 1$ and $r = 2$ largely overlap. For example, a HR of 0.8 or less for both OS and PFS would consist of a positive outcome in the adjuvant colon cancer setting regardless of r.

To empirically validate the assumption on the historical relationship, one may compare $\hat{\Delta}_{OS}$ to the 95% confidence interval of $\hat{\Delta}_P$, $\left(\hat{\Delta}_P - 1.96\sqrt{Var(\hat{\Delta}_P)}, \hat{\Delta}_P + 1.96\sqrt{Var(\hat{\Delta}_P)}\right)$. If it falls into the confidence interval, a positive conclusion

may be made with strong confidence. If it doesn't but $\hat{\Delta}_{OS}$ is less than $\hat{\Delta}_P$ (i.e., observed OS effect is stronger than estimated from PFS data), an even stronger conclusion may be made. Otherwise, the trial may not be considered positive at the 2.5% significance level even if the outcome is in the rejection region. (The outcome from the hypothetical trial passes this test under both disease settings.) Certainly, one may use other empirical approaches as appropriate to validate the assumption.

3 Event Size Calculations

In this section, we provide event size calculations for trials intending to use $\hat{Z}_J$ (7) for demonstrating a statistically significant OS effect at a one-sided type I error rate of α_J and type II error rate of β_J based on historical estimate of $(\hat{\gamma}, \hat{\sigma}_\gamma)$. For simplicity, the trials have two arms and apply a 1:1 randomization (minor modifications are needed otherwise.) Denote Δ_{PFS} as the PFS effect of interest and Δ_{OS} as the corresponding OS effect which is expected to be $\hat{\gamma}\Delta_{PFS}$. Suppose that the relative number of events for the PFS endpoint (N_{PFS}) and OS endpoint (N_{OS}) at the time of primary analysis (for application of accelerated approval) is relatively fixed at λ such that $N_{PFS} \approx \lambda N_{OS}$ ($\lambda > 1$). A ballpark estimate of λ for a typical Phase III oncology trial can be easily obtained upfront once a follow-up strategy is decided. For ease of illustration, we will keep λ at 4 in the following examples. This by no means implies that this is an optimal estimate for the respective disease settings we will consider. The sample sizes are inherent in $Var(\hat{\Delta}_J)$, which has to satisfy the following constraint

$$\Delta_{OS}^2 = (Z_{\alpha_J} + Z_{\beta_J})^2 \left[Var(\hat{\Delta}_J)\right] \tag{8}$$

or equivalently

$$\hat{\gamma}^2\Delta_{PFS}^2 = (Z_{\alpha_J} + Z_{\beta_J})^2 \left[w^2\hat{\sigma}_{OS}^2 + (1-w)^2\mathrm{Var}(\hat{\Delta}_P) + 2w(1-w)\hat{\gamma}\rho\,\hat{\sigma}_{OS}\hat{\sigma}_{PFS}\right] \tag{9}$$

where $\hat{\sigma}_{OS}^2 = 4/N_{OS}$ and $\hat{\sigma}_{PFS}^2 = 4/N_{PFS}$ with $Var(\hat{\Delta}_P)$ from (2) and w from (5). N_{PFS} and N_{OS} are readily solved from (9) once $\hat{\gamma}$, $\hat{\sigma}_\gamma$, λ, ρ, and Δ_{PFS} are provided. See appendix for sample code.

Table 1 provides N_{PFS} and N_{OS} and study properties at $\alpha_J = 2.5\%$ and $\beta_J = 10\%$ in the metastatic colorectal cancer setting. To meet the implicit OS objective at the time of analysis in which PFS is the primary endpoint in a conventional Phase III oncology trial, some thought leaders argue for a smaller type I error rate than 2.5% (one-sided) to be set for the PFS endpoint and some argue for testing of super-superiority. To compare with these alternative approaches, Table 1 also reports study powers for testing the null hypothesis $\mathrm{HR_{PFS}} = 1$ at 0.0625% (Pow1: equivalent to requirement of two positive trials) and testing the null hypothesis $\mathrm{HR_{PFS}} = 0.80$ at $\alpha = 2.5\%$ (Pow2: super-superiority) based on N_{PFS} using the conventional log-rank

Table 1 Sample sizes and study properties based on the joint test-statistic ($\hat{Z}_J$ of 7) at α_J=2.5% (one-sided) and $\beta_J = 10\%$ (i.e., 90% power) for metastatic colorectal cancer with $(\hat{\gamma}, \hat{\sigma}_\gamma) = (0.54, 0.1)$

Design parameters				PFS endpoint			OS endpoint		
r	ρ	Δ_{PFS}	Δ_{OS}	N_{PFS}	Pow1	Pow2	N_{OS}	Pow3	Pow4
1	0.3	log(0.5)	log(0.69)	147	84%	81%	37	20%	87%
		log(0.6)	log(0.76)	270	83%	66%	68	20%	87%
	0.5	log(0.5)	log(0.69)	165	89%	86%	41	22%	88%
		log(0.6)	log(0.76)	305	89%	71%	76	22%	88%
2	0.3	log(0.5)	log(0.69)	568	100%	100%	142	60%	99%
		log(0.6)	log(0.76)	1045	100%	100%	261	60%	99%
	0.5	log(0.5)	log(0.69)	627	100%	100%	157	64%	99%
		log(0.6)	log(0.76)	1154	100%	100%	289	65%	99%

The event ratio (λ) is fixed at 4
Pow1: power for testing null HR_{PFS}=1 at α=0.0625% (equivalent to requirement of two positive trials) based on N_{PFS} using the conventional log-rank test
Pow2: power for testing null HR_{PFS}=0.8 at α=2.5% (super-superiority) based on N_{PFS} using the conventional log-rank test
Pow3: power for testing null HR_{OS}=1 at α=2.5% (regular superiority) based on N_{OS} using the conventional log-rank test
Pow4: power for testing null HR_{OS}=1 at α=50% (positive trend) based on N_{OS} using the conventional log-rank test

test. Also reported are powers for testing the null hypothesis $HR_{OS} = 1$ at $\alpha = 2.5\%$ (Pow3: regular superiority) and $HR_{OS} = 1$ at $\alpha = 50\%$ (Pow4: positive trend) based on N_{OS} using the conventional log-rank test. Two PFS effects are considered, $HR = 0.5$ and $HR = 0.6$. The corresponding estimated OS effects are $HR = 0.69$ and $HR = 0.76$, respectively, based on historical data under the two disease settings. The sample size estimates are extremely sensitive to $\hat{\sigma}_\gamma$; they are more than tripled when $\hat{\sigma}_\gamma$ is doubled. The estimates are also sensitive to the effect size of interest but less so to ρ, as one might be expected. To demonstrate a statistically significant PFS effect at the same type I/II error rates, the number of PFS events would be 88 for $\Delta_{PFS} = \log(0.5)$ and 162 for $\Delta_{PFS} = \log(0.6)$. Not surprisingly, as seen from the table, more PFS events are needed to have sufficient power to demonstrate an OS effect with the proposed OS estimator. On the other hand, if a conventional survival trial at the same type I/II error rates is considered, the number of OS events would be 305 for $\Delta_{OS} = \log(0.69)$ and 558 for $\Delta_{OS} = \log(0.76)$, which is much greater than N_{OS} in the table.

Table 2 provides the same analyses in the adjuvant colon cancer setting except that the null hypothesis for super-superiority is changed to $HR = 0.90$ for PFS to better reflect the difference in $(\hat{\gamma}, \hat{\sigma}_\gamma)$. Two PFS effects are considered, $HR = 0.6$ and $HR = 0.7$. The corresponding OS effects are $HR = 0.63$ and $HR = 0.73$, respectively. As a comparison, the number of PFS events would be 162 for $\Delta_{PFS} = \log(0.6)$

Table 2 Sample sizes and study properties based on the joint test-statistic ($\hat{Z}_J$ of 7) at α_J=2.5% (one-sided) and $\beta_J = 10\%$ (i.e., 90% power) for adjuvant colon cancer with $(\hat{\gamma}, \hat{\sigma}_\gamma) = (0.89, 0.061)$

Design parameters				PFS endpoint			OS endpoint		
r	ρ	Δ_{PFS}	Δ_{OS}	N_{PFS}	Pow1	Pow2	N_{OS}	Pow3	Pow4
1	0.3	log(0.6)	log(0.63)	172	55%	76%	43	33%	94%
		log(0.7)	log(0.73)	353	55%	66%	89	31%	93%
	0.5	log(0.6)	log(0.63)	194	63%	81%	49	36%	95%
		log(0.7)	log(0.73)	398	63%	71%	100	35%	94%
2	0.3	log(0.6)	log(0.63)	198	64%	81%	50	37%	95%
		log(0.7)	log(0.73)	405	64%	72%	102	35%	94%
	0.5	log(0.6)	log(0.63)	225	73%	86%	57	41%	96%
		log(0.7)	log(0.73)	461	73%	77%	116	39%	95%

The event ratio (λ) is fixed at 4
Pow1: power for testing null HR=1 at α=0.0625% (equivalent to requirement of two positive trials) based on N_{PFS} PFS events using the conventional log-rank test
Pow2: power for testing null HR=80% at α=2.5% (super-superiority) based on N_{PFS} PFS events using the conventional log-rank test
Pow3: power for testing null HR=1 at α=2.5% (regular superiority) based on N_{OS} OS events using the conventional log-rank test
Pow4: power for testing null HR=1 at α=50% (positive trend) based on N_{OS} OS events using the conventional log-rank test

and 330 for $\Delta_{PFS} = \log(0.7)$ in a conventional PFS trial while the number of OS events would be 197 for $\Delta_{OS} = \log(0.63)$ and 425 for $\Delta_{OS} = \log(0.73)$ in a conventional survival trial at the same type I/II error rates. The sample size estimates in this disease setting are less sensitive to $\hat{\sigma}_\gamma$ compared to the metastatic colorectal cancer setting. Because OS effect can be estimated with greater certainty from PFS effect, N_{PFS} only has a moderate increase as compared to a conventional PFS trial. This partly confirms the general belief that PFS is a good surrogate endpoint for OS for adjuvant colon cancer. But it also implies that a conventional PFS trial would still be underpowered for answering the elusive OS question even when PFS is a good surrogate endpoint.

It is clear from the two tables that, even though the number of events required with the design based on our approach is larger than the size of a conventional trial in which PFS is the primary endpoint, it has smaller sample size than a conventional survival trial. By incorporating historical and current trial data, it can address the elusive OS question that a conventional PFS trial cannot, no matter how good a surrogate endpoint PFS is. It is also clear from the two tables that it is impossible to set a common bar on the type I error rate or null effect in a conventional PFS trial to guard against a potential spurious outcome, despite good intentions. A clinically meaningful PFS effect in one disease setting (e.g., adjuvant colon cancer) may not be so in a different setting (e.g., metastatic colorectal cancer) no matter

how compelling it is. Conversely, some disease settings do not need to have a PFS effect as large as in other settings to indicate an improvement in OS. It is also clear from the two tables that the definition of a positive survival trend has a drastically different implication to study power. Just like for the PFS endpoint, it is impossible to set a common bar without being ultraconservative or arbitrary. From similar analyses under various alternative assumptions on key parameters such as historical relationship and relative number of events (results not presented here), the general pattern of the above findings holds.

In planning for a Phase III trial that applies $\hat{Z}_J$ (7) to the testing of the OS effect, we may use data from Phase II and other similar trials to estimate ρ for the sample size calculation. While it makes event size calculation easier by fixing λ upfront, there is room for improvement when the design is complicated or when accrual and follow-up is not straightforward. Nevertheless, the event size (9), assisted with sensitivity analyses, provides a starting point for trial planning.

4 Discussion

Regulatory approval of a drug application to market is ultimately decided by a trade-off between risk and benefit. Traditional chemotherapies for oncology are toxic; they are usually approved after successful demonstration of an unequivocal survival benefit over standard of care. Many new biologics or small molecule drugs under development are target-specific and are presumably less toxic, thanks to the phenomenal expansion of our knowledge in the molecular biology of cancer. This raises a natural question whether a conventional survival trial should be mandatory for their approval. Of all the viable alternative endpoints that have been considered, PFS has received the most attention. However, for majority of the disease settings, the PFS endpoint can't be considered a valid surrogate endpoint to survival and is not necessarily a direct measure of clinical benefit. A statistically significant outcome on PFS does not automatically equate to a positive trial, which, to the minimum, requires providing evidence of survival benefit with reasonable assurance by FDA's AA mechanism. More and more often, regulatory agencies face the difficult task of setting an appropriate bar for PFS effect size for accelerated approval.

To address this regulatory issue, we have proposed a synthesized approach that incorporates all relevant data within and outside of a trial (i.e., PFS and OS data as well as historical relationship between the OS effect size and the PFS effect size) into a simple joint test-statistic for explicitly testing the implicit OS hypothesis. An event size equation is also provided for trials intending to use the joint test-statistic for hypothesis testing. The proposed approach is applied to hypothetical Phase III trials in metastatic colorectal cancer and adjuvant colon cancer settings. It demonstrates that rational and objective bars for PFS effect size as well as for OS effect size can be set using our proposed approach. Further, a modified design of the conventional PFS trial can be prospectively planned to directly address the

regulatory issue under the same approach. It has to be emphasized that, while the proposed approach helps address a relevant regulatory and scientific issue associated with FDA's AA mechanism, it is never intended to replace a conventional survival trial should it be a requirement for regulatory approval. It also has to be emphasized that we have treated the regulatory issue as a statistical estimation and hypothesis testing problem, which is a conventional approach in the regulatory environment, as opposed to a statistical prediction problem as in [13, 14]. While this represents our best understanding of the regulatory policy, it is up to the policy makers to clarify.

Reliable estimation of the historical relationship plays a pivotal role in successful estimation of OS effect from PFS data. In the last decade, tremendous clinical data for various cancer diseases have been generated, with much more forthcoming. They provide a solid foundation for establishing the much-needed historical relationships. Regulatory agencies and scientific advisors implicitly incorporate such a relationship, or lack thereof, into the decision process for drug approval. However, different reviewers may use different historical data sources to come up with their own estimates. They may come up with different estimates even they have access to the same historical data. Besides, as pointed out by an anonymous referee, even when an appropriate meta-analysis is applied different estimates may be obtained from different methods (e.g., fixed effects regression, random effects regression, and mean estimation of actual effect size ratios). More importantly, it is unclear how to fully account for the uncertainties in historical relationship. We used the standard error from meta-analysis (after possible inflation) in our approach, which is consistent with [4]. This essentially treats the current trial as a typical or "average" trial among historical ones, as an anonymous referee has correctly pointed out. However, it is unclear whether this represents an accurate interpretation of the regulatory policy. An alternative Bayesian approach can be found in [21].

Propelled by scientific breakthroughs, the current regulatory requirement for oncology drug approval has created an unmet need for the statistical community. A lot of energy and effort has been spent on surrogacy validation in the last two decades. But numerous fruitful findings have not been sufficient to address the emerging regulatory issues. The approach we have proposed in this paper represents a step in the right direction. To clearly relay the key message, we purposely keep the technical details to minimum. Many related statistical issues (e.g., group sequential design in light of a different variance–covariance structure, comparison between the MLE and the weighted estimator, properties of the joint test-statistic) and practical issues (e.g., filing for accelerated approval after an interim analysis while the trial is ongoing, impact of accelerated approval on execution of Phase IV commitment in a separate trial, impact of cross-in on survival analysis) are left open for further investigation.

The issues inherent in the AA mechanism (and to some extent conditional approval under CHMP) are both statistical and regulatory. A collaborative effort among all stakeholders in drug development (drug developers, regulatory reviewers, and academic researchers) is greatly needed to help make accelerated approval a more data-driven decision process.

References

Code of Federal Regulations- 21 CFR 314 and 601. *Accelerated Approval Rule*, 1992.

Burzykowski T, Buyse M, Yothers G, Sakamoto J and Sargent D. Exploring and validating surrogate endpoints in colorectal cancer. *Lifetime Data Analysis* 2008; 14: 54-64.

Sargent D, Wieand S, Haller DG et al. Disease-free survival (DFS) vs. overall survival (OS) as a primary endpoint for adjuvant colon cancer studies: Individual patient data from 20, 898 patents on 18 randomized trials. *J Clin Oncol* 2005; 23: 8664-8670.

Tang PA, Bentzen SM, Chen EX, and Siu LL. Surrogate End Points for Median Overall Survival in Metastatic Colorectal Cancer: Literature-Based Analysis From 39 Randomized Controlled Trials of First-Line Chemotherapy. *Journal of Clinical Oncology* 2007; 25: 4562-4568.

Miksad RA, Zietemann V, Gothe R, et al. Progression-free-survival as a surrogate endpoint in advanced breast cancer. *International Journal of Technology Assessment in Health Care* 2008; 24: 371-383.

Piedbois P, and Croswell, JM. Surrogate endpoints for overall survival in advanced colorectal cancer: A clinician's perspective. *Statistical Methods in Medical Research* 2008; 17: 519-527.

Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann. Intern. Med.* 1996; 125: 605–613.

Buyse M, Molenberghs G, Burzykowski T, Renard D, and Geys H. Statistical validation of surrogate endpoints: problems and proposals. *Drug Inf. J.* 2000; 34: 447–454.

Buyse M, Molenberghs G, Burzykowski T, Renard D and Geys, H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000; 1: 49-67.

Chen C, Wang HW, and Snapinn S. Proportion of treatment effect (PTE) explained by a surrogate marker. *Stat. Med.* 2003; 22: 3449–3459.

Chakravarty A, and Sridhara R. Use of progression-free survival as a surrogate marker in oncology trials: some regulatory issues. *Statistical Methods in Medical Research* 2008; 17: 515–518.

Fleming TR, Rothmann MD, and Lu HL. Issues in using progression-free survival when evaluating oncology products. *Journal of Clinical Oncology* 2009; 27: 2874–2880.

Burzykowski T, and Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharm Stat* 2006; 5: 173-186.

Buyse M, Burzykowski T, Carroll K et al. Progression-free-survival is a surrogate for survival in advanced colorectal cancer. *Journal of Clinical Oncology* 2007; 25: 5218–5224.

FDA Guidance for the use of Bayesian statistics in medical device clinical trials (draft). Last accessed 16 October 2009, from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm.

The International Conference on Harmonisation (ICH) Guidance "Choice of Control Group in Clinical Trials" (ICH –E10), Accessed 16 October 2009, http://www.ich.org/cache/compo/276-254-1.html.

SAS/STAT® Software: Changes and Enhancements through Release 6.12, Cary, NC: SAS Institute Inc., 1997.

Burzykowski T, Molenberghs G, Buyse M. *The evaluation of surrogate endpoints*. Springer, 2005 (ISBN 0-387-20277-3).

Whitehead, A. *Meta-analysis of controlled clinical trials*. UK: Wiley, 2002.

Chen C, Sun L. Quantification of PFS effect for accelerated approval of oncology drugs. *Statistics in Biopharmaceutical Research* 2011, volume 3, 434–444.

Daniels MJ, Hughes, MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistical in Medicine* 1997, volume 16, 1965–1982.

Integrative Analysis of Multiple Cancer Prognosis Datasets Under the Heterogeneity Model

Jin Liu, Jian Huang, and Shuangge Ma

Abstract In cancer research, genomic studies have been extensively conducted, searching for markers associated with prognosis. Because of the "large d, small n" characteristic, results generated from the analysis of a single dataset can be unsatisfactory. Integrative analysis simultaneously analyzes multiple datasets and can be more effective than the analysis of single datasets and classic meta-analysis. In many existing integrative analyses, the homogeneity model has been assumed, which postulates that different datasets share the same set of markers. In practice, datasets may have been generated in studies that differ in patient selection criteria, profiling techniques, and many other aspects. Such differences may make the homogeneity model too restricted. Here we explore the heterogeneity model, which assumes that different datasets may have different sets of markers. With multiple cancer prognosis datasets, we adopt the AFT (accelerated failure time) models to describe survival. A weighted least squares approach is adopted for estimation. For marker selection, penalization-based methods are examined. These methods have intuitive formulations and can be computed using effective group

J. Liu
Department of Biostatistics, School of Public Health, Yale University,
New Haven, CT 06520, USA
e-mail: jin.liu.jl2329@yale.edu

J. Huang
Department of Statistics and Actuarial Science, University of Iowa,
2 W Jefferson St, Iowa City, IA 52246, USA
e-mail: jian-huang@uiowa.edu

S. Ma (✉)
Department of Biostatistics, School of Public Health, Yale University,
New Haven, CT 06520, USA

VA Cooperative Studies Program Coordinating Center, West Haven, CT, USA
e-mail: shuangge.ma@yale.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_21, © Springer Science+Business Media New York 2013

coordinate descent algorithms. Analysis of three lung cancer prognosis datasets with gene expression measurements demonstrates the merit of heterogeneity model and proposed methods.

1 Introduction

Genomic studies have been extensively conducted, searching for markers associated with the prognosis of cancer. Data generated in such studies have the "large d, small n" characteristic, with the number of genes profiled d much larger than sample size n. In addition, in whole-genome studies, only a subset of the profiled genes are expected to be associated with prognosis. Thus, the analysis of cancer prognosis data with genomic measurements demands regularized estimation and selection.

In practical data analysis, genomic markers identified from the analysis of single datasets are often unsatisfactory. Multiple factors contribute to the unsatisfactory performance, including the highly noisy nature of cancer genomic data, technical variations of profiling techniques and, more importantly, the small sample sizes of individual studies. Recent studies have shown that pooling and analyzing multiple studies may effectively increase sample size and improve properties of the identified markers (Guerra and Goldsterin 2009; Ma et al. 2009, and references therein). Multi-dataset methods include meta-analysis and integrative analysis methods. Integrative analysis pools and analyzes raw data from multiple studies and can be more informative than classic meta-analysis, which analyzes multiple studies separately and then pools summary statistics (lists of identified genes, p-values, effect sizes, etc.).

In studies such as Ma et al. (2011b), the homogeneity model has been assumed. Under this model, multiple datasets share the same set of markers. This model has also been adopted with cancer diagnosis studies and categorical responses (Ma et al. 2011a and references therein). In practical data analysis, when multiple datasets are generated in independent studies, heterogeneity (in patients' characteristics, technical aspects such as profiling protocols, etc) inevitably exists. Such heterogeneity may make the homogeneity model too restricted. In addition, data analyses in Ma et al. (2011a;b) show that for some of the identified genes, the magnitudes of estimated regression coefficients may vary significantly across datasets. It is possible that the very small regression coefficients are actually zero. Such an observation further suggests the necessity of relaxing the homogeneity model assumption.

In this study, we describe cancer survival using AFT (accelerated failure time) models. Compared with alternatives such as the Cox model, the AFT model has a significantly simpler objective function and lower computational cost, which is especially desirable with high-dimensional data. In addition, its regression coefficients may have more lucid interpretations. As an alternative to the homogeneity model, we consider the heterogeneity model. It includes the homogeneity model as a special case and can be more flexible. For marker selection, we adopt

penalization. The proposed penalization methods are intuitively reasonable and can be computationally realized using the group coordinate descent algorithms. This study complements the existing ones by conducting integrative analysis under the more flexible heterogeneity model and by adopting penalization methods tailored to this model.

2 Integrative Analysis of Cancer Prognosis Studies

2.1 Data and Model Settings

Assume M independent studies and n^m iid observations in study $m(=1,\ldots,M)$. The total sample size is $n=\sum_{m=1}^{M} n^m$. In study m, denote T^m as the logarithm (or another known monotone transformation) of failure time. Denote X^m as the length-d vector of gene expressions. Although gene expression data is used as an example in this study, it should be noted that the proposed methods are also applicable to studies with other types of genomic measurements. For simplicity of notation, assume that the same set of genes are measured in all M studies. For the ith subject, the AFT model assumes that

$$T_i^m = \beta_0^m + X_i^{m\prime}\beta^m + \epsilon_i^m, i=1,\ldots,n^m. \tag{1}$$

where β_0^m is the intercept, $\beta^m \in \mathbb{R}^d$ is the length-d vector of regression coefficients, and ϵ_i^m is the error term. When T_i^m is subject to right censoring, we observe $(Y_i^m, \delta_i^m, X_i^m)$, where $Y_i^m = \min\{T_i^m, C_i^m\}$, C_i^m is the logarithm of censoring time, and $\delta_i^m = I\{T_i^m \le C_i^m\}$ is the event indicator.

When the distribution of ϵ_i^m is known, the parametric likelihood function can be easily constructed. Here we consider the more flexible case where this distribution is unknown. In the literature, multiple estimation approaches have been developed, including, for example, the Buckley-James and rank-based approaches (Buckley and James 1979; Jin et al. 2003). In this study, we adopt the weighted least squares estimator (Stute 1996), which to the best of our knowledge, has the lowest computational cost. This property is especially desirable with high-dimensional data.

Let $\hat{F}^m$ be the Kaplan–Meier estimator of the distribution function F^m of T^m. $\hat{F}^m(y) = \sum_{i=1}^{n^m} \omega_i^m I\{Y_{(i)}^m \le y\}$, where ω_i^ms are the jumps in the Kaplan–Meier estimator and can be computed as

$$\omega_1^m = \frac{\delta_{(1)}^m}{n^m},\ \omega_i^m = \frac{\delta_{(i)}^m}{n^m - i + 1}\prod_{j=1}^{i-1}\left(\frac{n^m - j}{n^m - j + 1}\right)^{\delta_{(j)}^m},\ i=2,\ldots,n^m.$$

Here $Y_{(1)}^m \le \cdots \le Y_{(n^m)}^m$ are the order statistics of Y_i^ms, and $\delta_{(1)}^m,\ldots,\delta_{(n^m)}^m$ are the associated event indicators. Similarly, let $X_{(1)}^m,\ldots,X_{(n^m)}^m$ be the associated gene

expressions of the ordered Y_i^ms. Stute (1996) proposed the weighted least squares estimator $(\hat{\beta}_0^m, \hat{\beta}^m)$ that minimizes

$$\frac{1}{2}\sum_{i=1}^{n^m} \omega_i^m (Y_{(i)}^m - \beta_0^m - X_{(i)}^{m\prime}\beta^m)^2. \tag{2}$$

We center $X_{(i)}^m$ and $Y_{(i)}^m$ using their ω_i^m-weighted means, respectively. Define

$$\bar{X}_w^m = \sum_{i=1}^{n^m} \omega_i^m X_{(i)}^m / \sum_{i=1}^{n^m} \omega_i^m, \bar{Y}_w^m = \sum_{i=1}^{n^m} \omega_i^m Y_{(i)}^m / \sum_{i=1}^{n^m} \omega_i^m.$$

Let $X_{\omega(i)}^m = \sqrt{\omega_i^m}(X_{(i)}^m - \bar{X}_w^m)$ and $Y_{\omega(i)}^m = \sqrt{\omega_i^m}(Y_{(i)}^m - \bar{Y}_w^m)$, respectively. With the weighted centered values, the intercept is zero. The weighted least squares objective function can be written as

$$L^m(\beta^m) = \frac{1}{2}\sum_{i=1}^{n^m} (Y_{\omega(i)}^m - X_{\omega(i)}^{m\prime}\beta^m)^2. \tag{3}$$

Denote $Y^m = (Y_{\omega(1)}^m, \ldots, Y_{\omega(n^m)}^m)'$ and $X^m = (X_{\omega(1)}^m, \ldots, X_{\omega(n^m)}^m)'$. Further denote $Y = (Y^{1\prime}, \ldots, Y^{M\prime})'$, $X = \text{diag}(X^1, \ldots, X^M)$, and $\beta = (\beta^{1\prime}, \ldots, \beta^{M\prime})'$.

Consider the overall objective function $L(\beta) = \frac{1}{n}\sum_{m=1}^M L^m(\beta^m)$. With this objective function, larger datasets have more contributions, which is intuitively reasonable. When desirable, normalization by sample size can be applied.

2.2 Homogeneity Model and Penalized Selection

In Huang et al. (2012) and Ma et al. (2011a;b), the homogeneity model is adopted to describe the genomic basis of M datasets. Denote β_j^m as the jth component of β^m. Then $\beta_j = (\beta_j^1, \ldots, \beta_j^M)'$ is the length-M vector of regression coefficients representing the effects of gene j in M studies. Under the homogeneity model, for any $j(= 1, \ldots, d)$,

$$I(\beta_j^1 = 0) = \ldots = I(\beta_j^M = 0).$$

That is, if a gene is identified as associated with prognosis in one dataset, it is identified in all of the M datasets. Thus, the M datasets have the same sparsity structure. This is a sensible model when multiple datasets have been generated under the same protocol. With multiple datasets generated independently, if the analysis of individual datasets and examination of the protocols suggest a high degree of similarity, then this model can be adopted.

For marker selection, Ma et al. (2011b) adopts penalization and proposes using the group MCP (gMCP) approach, where the estimate is defined as

$$\hat{\beta} = argmin\left\{L(\beta) + P_{gMCP}(\beta)\right\},$$

with

$$P_{gMCP}(\beta) = \sum_{j=1}^{d} \rho(||\beta_j||_{\Sigma_j}; \sqrt{d_j}\lambda_1, \gamma). \tag{4}$$

$\rho(t;\lambda,\gamma) = \lambda \int_0^{|t|} \left(1 - \frac{x}{\lambda\gamma}\right)_+ dx$ is the MCP penalty (Zhang 2010). $||\beta_j||_{\Sigma_j} = ||\Sigma_j^{1/2}\beta_j||_2$, $||\cdot||_2$ is the L_2 norm, $\Sigma_j = n^{-1}X[\cdot,j]'X[\cdot,j]$, and $X[\cdot,j]$ is the $n \times d_j$ submatrix of X that corresponds to β_j. In (4), d_j is the size of the coefficient group corresponding to gene j. When the M datasets have exactly matched gene sets, $d_j \equiv M$. We keep d_j so that this formulation can accommodate partially matched gene sets. When gene j is not measured in dataset k, we take the convention $\beta_j^k \equiv 0$. $\lambda_1 > 0$ is the tuning parameter, and $\gamma > 0$ is the regularization parameter (Zhang 2010).

Penalty function (4) has been motivated by the following considerations. In this analysis, genes are the functional units. The overall penalty is the sum over d individual penalties, with one for each gene. For gene selection, the MCP penalization is adopted. In single-dataset analysis, MCP has been shown to have performance comparable to or better than some of the alternative penalties. For a specific gene, its effects in the M studies are represented by a "group" of M regression coefficients. Under the homogeneity model, the M studies are expected to identify the same set of genes. Thus, within a group, no further selection is needed, and so the L_2 norm is adopted. Note that here we adopt the $||\cdot||_{\Sigma_j}$ norm, which rescales the regression coefficient vector by the covariance matrix Σ_j, so that the computation can be less ad hoc.

3 Heterogeneity Model and Penalized Selection

3.1 Heterogeneity Model

When multiple datasets are generated in independent studies, heterogeneity inevitably exists (Knudsen 2006). The degree of heterogeneity depends on the differences in study protocols, profiling techniques, and many other factors. In cancer prognosis studies, the effort to unify the sets of identified markers across independent studies has not been very successful (Cheang et al. 2008; Knudsen 2006). This can also be partly seen from the data analysis in Ma et al. (2011b). Such observations raise the question whether the homogeneity model is too restricted and motivates the heterogeneity model. Under the heterogeneity model, one gene can be associated with prognosis in some studies but not others. This model includes the homogeneity model as a special case and can be more flexible.

In addition, there are scenarios under which the homogeneity model is conceptually not sensible, but the heterogeneity model is. The first is where different

studies are on different types of cancers (Ma et al. 2009). On the one hand, different cancers have different prognosis patterns and different sets of markers. On the other hand, multiple pathways, such as apoptosis, DNA repair, cell cycle, and signaling, are associated with the prognosis of multiple cancers. The second scenario is the analysis of different subtypes of the same cancer. Different subtypes have different risks of occurrence and progression, and it is not sensible to reinforce the same genomic basis. The third scenario is where subjects in different studies have different demographic measurements, clinical risk factors, environmental exposures, and treatment regimens. For genes not intervened with those "additional" variables, their importance remains consistent across multiple studies. However, for other genes, they may be important in some studies but not others.

3.2 Penalized Marker Selection

Under the heterogeneity model, the model and regression coefficients have two dimensions. The first is the gene dimension as in other marker selection studies. The second, which is unique to integrative analysis, is the study (dataset) dimension. In marker selection, we need to determine *whether a gene is associated with prognosis in any study at all* as well as *in which studies it is associated with prognosis*. Such an objective demands two-way selection.

3.2.1 A Two-Step Approach

Since integrative analysis under the heterogeneity model calls for two-way selection, a natural strategy is to achieve the selection in two steps, with one step for each way. The first step is to determine whether a gene is associated with prognosis in any study. As $\beta_j = (\beta_j^1, \ldots, \beta_j^M)'$ represent the effects of gene j $(= 1, \ldots, d)$ in M studies, this step of selection amounts to determining whether $\|\beta_j\|_2 = 0$. We propose achieving this step of selection using the *gMCP penalization approach*. With genes selected in the first step (i.e., $\{j : ||\beta_j||_2 \neq 0\}$), in the second step, we determine which prognosis responses (studies) they are associated with. For this step, we propose applying the *MCP approach* to each dataset separately. This step conducts standard single-dataset analysis. Note that although both steps employ the MCP penalties, they may have different tuning and regularization parameters.

3.2.2 Composite Penalization

In the analysis of a single dataset, when there is a grouping structure among covariates, two-way selection can be achieved using composite penalization. The idea is to use a group penalty for variable selection at the group level and a second penalty for variable selection at the within-group level. The composite of the two penalties will then be able to conduct variable selection at both levels.

In integrative analysis with multiple datasets, we adopt a similar strategy. First consider the *1-norm gMCP* approach, where the penalty takes the form

$$P_{1-norm\ gMCP}(\beta) = \sum_{j=1}^{d} \rho(\|\beta_j\|_1; \sqrt{d_j}\lambda, \gamma). \tag{5}$$

$\|\beta_j\|_1 = \sum_{m=1}^{M} |\beta_j^m|$ is the L_1 norm of β_j, which can also be viewed as a Lasso penalty. Other notations have similar implications as with gMCP. With this composite penalty, the outer penalty has a gMCP form. In integrative analysis, it conducts selection at the gene level. The inner penalty is Lasso. For a gene with nonzero effects, it identifies in which study(ies) the gene is associated with prognosis.

In (5), the Lasso penalty is adopted mainly because of its computational simplicity. In single-dataset analysis, it has been shown that MCP can have better properties (for example, more accurate selection) than Lasso. Motivated by such a consideration, we propose the *composite MCP (cMCP)* approach, where the penalty takes the form

$$P_{cMCP}(\beta) = \sum_{j=1}^{d} \rho\left(\sum_{m=1}^{M} \rho(\beta_j^m; \lambda_2, b); \lambda_1, a\right). \tag{6}$$

Here λ_1, λ_2 are the tuning parameters, and a, b are the regularization parameters.

Computational algorithm for cMCP . Below we describe the computational algorithm for cMCP. The two-step and 1-norm gMCP estimates can be computed in a similar manner.

Consider the group coordinate descent algorithm. This algorithm is iterative and optimizes over the regression coefficients of one gene at a time. It cycles through all genes, and the overall iteration is repeated multiple times until convergence. Here the key is update of estimates for a single group (gene). Unfortunately, the cMCP approach does not have a simple form for updating individual groups. To tackle this problem, we adopt an approximation approach. Consider update with the jth group. By taking the first order Taylor series approximation about β_j and evaluating at $\tilde{\beta}_j$ (the current estimate), the penalty as a function of β_j^k is approximately proportional to $\tilde{\lambda}_{jk}|\beta_j^k|$ where

$$\tilde{\lambda}_{jk} = \rho'\left(\sum_{m=1}^{M} \rho(|\tilde{\beta}_j^m|; \lambda_2, b); \lambda_1, a\right) \rho'(|\tilde{\beta}_j^k|; \lambda_2, b). \tag{7}$$

For update with each β_j^k, we have an explicit solution:

$$\hat{\beta}_j^k = f_{cMCP}(z; \lambda) = S_1(z, \lambda), \tag{8}$$

with $S_1(z, \lambda) = sgn(z)(|z| - \lambda)_+$, and z and λ to be defined below.

Consider the following algorithm. With fixed tuning and regularization parameters,

1. Initialize $s = 0$, the estimate $\beta^{(0)} = (\beta_1^{(0)\prime}, \ldots, \beta_d^{(0)\prime})' = (0, \ldots, 0)'$, and the vector of residuals $r = Y - X\beta^{(0)}$;
2. For $j = 1, \ldots, d$,
   (a) Calculate $\tilde{\lambda}_{jk}$ according to expression (7).
   (b) Calculate $z_j = n^{-1}X[\cdot, j]'r + \beta_j^{(s)}$. $X[\cdot, j]$ is the $n \times d_j$ submatrix of X that corresponds to β_j.
   (c) For $k = 1, \ldots, M$, update $\beta_j^{k(s+1)} \leftarrow f_{cMCP}(z_j^k; \tilde{\lambda}_{jk})$, where z_j^k is the kth element of z_j.
   (d) Update $r \leftarrow r - X[\cdot, j](\beta_j^{(s+1)} - \beta_j^{(s)})$.

   Update $s \leftarrow s + 1$.
3. Repeat Step 2 until convergence.

We use the L_2 norm of the difference between two consecutive estimates smaller than 0.001 as the convergence criterion. Convergence is achieved for the lung cancer datasets within twenty iterations. For the proposed methods, in the objective function, the first term is continuously differentiable and regular in the sense of Tseng (2001). The penalty term is separable. Thus the coordinate descent estimate converges to a coordinate-wise minimum of the first term, which is also a stationary point. Our limited experience suggests that the proposed computational algorithms are affordable. Among the three approaches, cMCP has the highest computational cost. With fixed tunings, the analysis of the lung cancer datasets (Sect. 4) takes about 40 s using a regular desktop PC.

3.2.3 Tuning Parameter Selection

The proposed methods involve the following tuning/regularization parameters: two-step approach: (λ, γ) for gMCP and possibly different (λ, γ) for MCP, 1-norm gMCP: (λ, γ), and cMCP: $(\lambda_1, \lambda_2, a, b)$.

Properties of the estimates are jointly determined by the tuning/regularization parameters. Generally speaking, smaller values of a and b (γ in MCP and gMCP) are better at retaining the unbiasedness of the MCP penalty for large coefficients, but they also have the risk of creating objective functions with a nonconvexity problem that are difficult to optimize and yield solutions that are discontinuous with respect to λ_1 and λ_2 (λ). It is therefore advisable to choose values of a and b (γ) that are big enough to avoid this problem but not too big. As suggested in Breheny and Huang (2011) and Zhang (2010), we have experimented with a few values for a and b (γ), particularly including 1.8, 3, 6, and 10.

In our numerical study, we select tuning parameters via V-fold cross validation with $V = 5$. Our limited unpublished simulation suggests that $a = 6$, $b = 6$ and $\gamma = 6$ lead to the best performance. We note that such a result does not indicate the

universal superiority of those values. In practice, searching over multiple possible values is still needed. With λ (λ_1, λ_2), one may expect that its value cannot go down to very small values since there are regions not locally convex (Breheny and Huang 2009; 2011). The criteria over non-locally convex regions may go up and down. To avoid the unexpectedness of such regions, we select λ (λ_1, λ_2) where the criterion first goes up (see Breheny and Huang 2011 for related discussions).

4 Analysis of Lung Cancer Prognosis Studies

Lung cancer is the leading cause of death from cancer for both men and women in the USA and in most other parts of the world. Non-small-cell lung cancer (NSCLC) is the most common cause of lung cancer death, accounting for up to 85% of such deaths (Tsuboi et al. 2007). Gene profiling studies have been extensively conducted on lung cancer, searching for markers associated with prognosis. Three studies are described in Xie et al. (2011). The UM (University of Michigan Cancer Center) study had a total of 175 patients, among whom 102 died during follow-up. The median follow-up was 53 months. The HLM (Moffitt Cancer Center) study had a total of 79 subjects, among whom 60 died during follow-up. The median follow-up was 39 months. The CAN/DF (Dana-Farber Cancer Institute) study had a total of 82 patients, among whom 35 died during follow-up. The median follow-up was 51 months. We refer to Xie et al. (2011) and references therein for more details on study designs, subjects' characteristics, and profiling protocols; 22,283 genes were profiled in all three studies.

In previous studies such as Xie et al. (2011), the three datasets were combined and analyzed. Such a strategy corresponds to a special case of the homogeneity model in the present study. As the three datasets were generated in three independent studies, heterogeneity is expected to exist across datasets. This can be partly seen from the summary survival data and profiling protocols. Here we assume the heterogeneity model and analyze using the two-step method (Table 1), 1-norm gMCP (Table 2), and cMCP (Table 3). Note that with all methods, the small magnitudes of regression coefficients are caused by the "clustered" log survival times. The estimates suggest that different datasets may have different prognosis-associated genes. This partly explains why published studies have failed to unify the identified markers across different lung cancer prognosis datasets. As described in Sect. 1, multiple factors may contribute to this heterogeneity. Without having access to all the experiment details, we are not able to determine the exact cause of heterogeneity. Although there are overlaps, different approaches identify different sets of genes. Such an observation is not surprising and has been made in published studies such as Ma et al. (2011b).

To provide a more comprehensive description of the three datasets and various methods, we also conduct the evaluation of prediction performance. Although in principle marker identification and prediction are two distinct objectives, evaluation of prediction performance can be informative for marker identification. In particular,

Table 1 Two-step method: identified genes and their estimates.

Probe	Gene	UM	HLM	CAN/DF
200041_s_at	DDX39B	0.004		0.027
200642_at	SOD1		0.016	0.016
200650_s_at	LDHA	−0.007	0.044	
200694_s_at	DDX24		−0.031	
200747_s_at	NUMA1	−0.016	−0.022	
200772_x_at	PTMA	−0.029	0.034	−0.025
201021_s_at	DSTN		−0.050	
201033_x_at	RPLP0	−0.006		
201508_at	IGFBP4	0.0004	−0.050	
201523_x_at	UBE2N	0.017		
201568_at	UQCRQ	0.016	0.000	0.002
201789_at	Hs.59719		−0.058	
201875_s_at	MPZL1	−0.015		0.001
202081_at	IER2	0.0002	−0.026	
202162_s_at	CNOT8	0.006	0.017	0.001
202176_at	ERCC3	0.006		

Table 2 1-norm gMCP: identified genes and their estimates.

Probe	Gene	UM	HLM	CAN/DF
200041_s_at	DDX39B			0.005
200633_at	UBB		−0.001	
200642_at	SOD1		0.0004	2.8E-05
200674_s_at	RPL32		0.002	
200693_at	YWHAQ		−0.003	
200694_s_at	DDX24		−0.002	
200724_at	RPL10	0.0005		
200772_x_at	PTMA	−0.0002		−0.002
200804_at	TMBIM6	−0.001		
200972_at	TSPAN3	−0.003		
200973_s_at	TSPAN3			0.003
201021_s_at	DSTN		−0.016	
201033_x_at	RPLP0		−0.0005	
201173_x_at	NUDC		0.003	
201201_at	CSTB		0.005	
201508_at	IGFBP4		−0.001	
201611_s_at	ICMT			−0.001
201645_at	TNC	0.003		
201729_s_at	KIAA0100	5.0E-05		
201789_at	Hs.59719		−0.012	
202081_at	IER2		−0.002	
202146_at	IFRD1			−0.0001
202176_at	ERCC3		−0.002	
202183_s_at	KIF22		0.002	
202413_s_at	USP1			−0.001

Table 3 cMCP: identified genes and their estimates.

Probe	Gene	UM	HLM	CAN/DF
200041_s_at	DDX39B			0.005
200633_at	UBB		−0.001	
200674_s_at	RPL32		0.002	
200693_at	YWHAQ		−0.005	
200772_x_at	PTMA			−0.001
200972_at	TSPAN3	−0.002		
200973_s_at	TSPAN3			0.002
201021_s_at	DSTN		−0.017	
201033_x_at	RPLP0		−0.0003	
201173_x_at	NUDC		0.001	
201201_at	CSTB		0.004	
201645_at	TNC	0.001		
201789_at	Hs.59719		−0.011	
202176_at	ERCC3		−0.0003	
202183_s_at	KIF22		0.001	
202413_s_at	USP1			−0.0002

if prediction is more accurate, then the identified markers are expected to be more meaningful. For prediction evaluation, we adopt a random sampling approach as in Ma et al. (2009). More specifically, we generate training sets and corresponding testing sets by random splitting data (with sizes 3:1). Estimates are generated using the training sets only. We then make prediction for subjects in the testing sets. We dichotomize the predicted linear risk scores $X\hat{\beta}$ at the median, create two risk groups, and compute the logrank statistic, which measures the difference in survival between the two groups. To avoid extreme splits, this procedure is repeated 100 times. The average logrank statistics are calculated as 2.17 (two-step), 4.77 (1-norm gMCP), and 3.70 (cMCP). 1-norm gMCP is the only approach that can separate subjects into groups with significantly different survival risks (p-value = 0.029). Based on this prediction evaluation, genes and estimates presented in Table 2 are suggested as the final results for these three datasets.

5 Discussion

In cancer genomic research, multi-dataset analysis provides an effective way to overcome certain drawbacks of single-dataset analysis. In most published studies, it has been reinforced that multiple datasets share the same set of prognosis-associated genes, that is, the homogeneity model. In this study, for multiple cancer prognosis datasets, we consider the heterogeneity model, which includes the homogeneity model as a special case and can be less restricted. This model may provide a way to explain the failure to unify cancer prognosis markers across independent studies (Knudsen 2006; Cheang et al. 2008). Under the heterogeneity model, we propose

three penalization methods for marker identification. Such methods are intuitively reasonable and computationally feasible. Analysis of three lung cancer studies demonstrates the practical feasibility of proposed methods.

Under the heterogeneity model, marker selection needs to be conducted in two dimensions. Methods beyond penalization, for example thresholding and boosting, may also be able to achieve such selection. Comprehensive investigation and comparison of different approaches are beyond the scope of this article. The proposed methods are based on the MCP penalty, which has been shown to have satisfactory performance in single-dataset analysis. We suspect that it is possible to develop similar approaches based on, for example, bridge and SCAD penalties. As in single-dataset analysis there is no evidence that such penalties are superior to MCP, such a development is not pursued.

Acknowledgments We would like to thank the attendants of the ICSA 2012 Applied Statistics Symposium for valuable comments and organizers of the proceedings. The authors have been supported by the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development, awards CA142774, CA165923, CA152301 from NIH, DMS0904181 from NSF, and award 2012LD001 from National Bureau of Statistics of China.

References

Breheny, P. and Huang, J. (2009) Penalized methods for bi-level variable selection. Statistics and Its Interface. 2: 369–380.

Breheny, P. and Huang, J. (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Annals of Applied Statistics. 5: 232–253.

Buckley, I.V. and James, I. (1979) Linear regression with censored data. Biometrika. 66: 429–436.

Cheang, M.C.U., van de Rijn, M. and Nielsen, T.O. (2008) Gene expression profiling of breast cancer. Annual Reviews of Pathology: Mechanisms of Disease. 3, 67–97.

Guerra, R. and Goldsterin, D.R. (2009). Meta-Analysis and Combining Information in Genetics and Genomics. Chapman and Hall/CRC, 1st edition.

Huang, Y., Huang, J., Shia, B.C. and Ma, S. (2012) Identification of cancer genomic markers via integrative sparse boosting. Biostatistics. 13: 509–522.

Jin, Z., Lin, D.Y., Wei, L.J. and Ying, Z. (2003) Rank-based inference for the accelerated failure time model. Biometrika. 90: 341–353.

Knudsen, S. (2006) Cancer Diagnostics with DNA Microarrays. Wiley.

Ma, S., Huang, J. and Moran, M. (2009) Identification of genes associated with multiple cancers via integrative analysis. BMC Genomics, 10: 535.

Ma, S., Huang, J. and Song, X. (2011a) Integrative analysis and variable selection with multiple high-dimensional datasets. Biostatistics, 12: 763–775.

Ma, S., Huang, J., Wei, F., Xie, Y. and Fang, K. (2011b) Integrative analysis of multiple cancer prognosis studies with gene expression measurements. Statistics in Medicine. 30: 3361–3371.

Stute, W. (1996) Distributional convergence under random censorship when covariables are present. Scandinavian Journal of Statistics. 23: 461–471.

Tseng, P. (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of Optimization Theory and Applications. 109: 475–494.

Tsuboi, M., Ohira, T., Saji, H., Hiyajima, K., Kajiwara, N., Uchida, O. et al. (2007) The present status of postoperative adjuvant chemotherapy for completely resected non-small cell lung cancer. Ann Thorac Cardiovasc Surg. 13:73–77.

Xie, Y., Xiao, G., Coombes, K., Behrens, C., Solis, L., Raso, G., Girard, L., Erickson, H., Roth, J., Heymach, J., Moran, C., Danenberg, K., Minna, J. and Wistuba, I. (2011) Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non small cell lung cancer patients. Clin Cancer Res, 17(17): 5705–5714.

Zhang, C.H. (2010) Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics. 38: 894–942.

Part VI
Safety and Risk Analysis

On Analysis of Low Incidence Adverse Events in Clinical Trials

G. Frank Liu

Abstract In drug or vaccine development, some adverse events (AEs) of interest may occur infrequently. Because of their clinical importance, those AEs may be studied in a clinical trial with large sample size, long-term follow-up, or in meta-analysis of combined data from multiple trials. The conventional summary and analysis methods based on frequency of first occurrence and comparing the proportion difference between treatment groups may not be the best approach because (1) the drug exposure information is not considered in the frequency summary and analysis and (2) any recurrence of an event in the long-term follow-up is not accounted for. When recurrence events are considered, issues on the analysis such as intra-subject correlation among the recurrence events, over-dispersion, and zero inflation may need to be considered. In this paper, we review several approaches for summary and analysis of safety data in these settings. Considerations are given on the assumptions of the risk function, adjustment for differential follow-up, and handling of over-dispersion and excessive zero for low incidence events. Applications to drug and vaccine clinical trials will be used for demonstration.

1 Introduction

Safety analysis is critically important for clinical trials in testing new drug, biologic, or vaccine products. Sponsors are required to assess the safety profile of a new product by quantifying the incidence of potential adverse events (AEs) and comparing that with an active and/or placebo control. An overview of collecting and analysis of safety data in clinical trials can be found in Chuang-Stein (1998).

G.F. Liu, Ph.D. (✉)
Merck Research Laboratories, North Wales, PA 19454, USA
e-mail: guanghan_frank_liu@merck.com

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_22, © Springer Science+Business Media New York 2013

In clinical trials, an AE can be any unfavorable and unintended sign, symptom, or disease associated with the use of the study products. For some AEs of special interest, the incidence rates (i.e., the probability of observing an AE in a given patient-time unit) can be low. Because of clinical importance, these AEs may be studied in clinical trial setting before the product is approved. Due to low incidence, a study with large sample size and long-term follow-up, or a meta-analysis which combines multiple trials may be used. The conventional summary and analysis methods based on frequency of first occurrence and comparing the proportion difference between treatment groups may not be the best approach because (1) the drug exposure information is not considered in the frequency summary and analysis, which can be issues for meta-analysis or studies with differential follow-up and (2) any recurrence of an event in the long-term follow-up is not accounted for.

For low incidence events, the AE count observed in long-term follow-up can be approximated by a Poisson process or a sum of sequential binary outcomes. Statistical models such as Poisson regression is often used to estimate and compare the incidence rates. Several issues on analyzing these AEs include intra-subject correlation among the sequential outcomes, over-dispersion, and zero inflation. In this paper, we review several approaches for summary and analysis of safety data under these settings. Considerations are given on the assumptions of the risk function, adjustment for differential follow-up, and handling of over-dispersion and excessive zero for low incidence events. Both exposure-adjusted incidence rate and estimates from over-dispersion and zero inflation statistical models for counts data will be discussed. Applications to drug and vaccine clinical trials will be used for demonstration.

2 Common Measures for Analysis of Adverse Events

2.1 Crude Percentage

Adverse events are generally tabulated and listed for randomized clinical trials. One of the most common measures for tabulation is the crude percentage of AEs. It is calculated as

$$\text{Crude}\,\% = \text{n}/\text{N} \times 100\%, \tag{1}$$

where N is the number of patients who took at least one dose of study product and n is the number of patients with the specific AE observed during the follow-up.

The crude percentage provides a simple and easily understandable measure. The unit for the calculation is patient. It can be interpreted as an estimate of the cumulative probability of having an event at the end of the follow-up period assuming that all subjects in the study are followed to the end. It may underestimate the cumulative probability if some subjects discontinue from the study early. If the

proportions of discontinuation or the follow-up differ between treatment groups, the treatment comparison based on the crude percentage could be biased.

By definition, the crude percentage does not take into account for duration of follow-up or any recurrent event observed for a patient, which may be fine for short-term studies in which almost all patients may be followed for the duration of the study. In long-term studies, however, it may not be an optimal measure because the follow-up time may be different due to dropouts.

2.2 Exposure-Adjusted Incidence

To take account for differential follow-up, an incidence rate is often used which is defined as the number of subjects with a specific AE divided by the total exposure time (Liu et al. 2006; Siddiqui 2009). It is also called the exposure-adjusted incidence rate (EAIR), specifically,

$$\text{EAIR} = \text{n} / \sum \text{t}_\text{i} \times 100, \tag{2}$$

where n is the number of patients with the AE and t_i is the exposure time (time-to-first AE or follow-up time if the patient has no AE) for patient i. The EAIR is interpreted as number of AEs per 100 patient-time (e.g., patient-year).

The definition of EAIR is based on the assumption that the occurrence of a specific AE follows an independent Poisson process; therefore, event occurs with a constant rate over time. While this may not be appropriate for AEs that usually occur early in the study (e.g., allergic reactions) or events whose hazard rates increase or decrease over time, it can be a reasonable approximation for low incidence AEs in long-term studies.

2.3 Exposure-Adjusted Overall Incidence

In the definition of EAIR, only the first event and time-to-first event are considered. When recurrent events are observed in long-term studies, an overall incidence rate is another alternative measure which accounts for all recurrent events and follow-up.

In general, there is no simple explicit formula to calculate the overall incidence rate. Statistical models such as a Poisson regression may be used to estimate the overall incidence rate and to compare the rates between treatment groups. In clinical applications, the recurrent events may be correlated unless the events follow a Poisson process; therefore, the counts may have over-dispersion or zero inflation problems. We will discuss these statistical models and treatment comparisons in the next section.

3 Analysis Methods for Low Incidence AEs

3.1 Comparing Crude Percentage or EAIRs

For comparing percentages or EAIRs of a specific AE between 2 treatment groups, several metrics may be considered, including risk difference ($p_1 - p_2$) and relative risk (p_1/p_2). For percentages of AEs, odds ratio $(p_1(1-p_2)/(1-p_1)/p_2)$ is another possible metric. The choice of a metric is somewhat arbitrary but may provide quite different interpretations. For example, if the incidence of a particular AE is 1% on placebo compared to 2% on test drug, the relative risk increase is 100%, while the risk difference is 1 percentage point. While both are legitimate summaries, the risk difference is often preferred because of its simplicity and can always be calculated. For rare AEs, the ratio and odds ratio may not be defined if there are no events in the control treatment group. Therefore, we will focus our discussions on the risk difference in this paper.

For proportion difference, it is assumed that $n_i \sim \text{Binomial}(N_i, p_i)$, $i = 1,2$. The point estimate for $\theta = p_1 - p_2$ is $\hat{\theta} = n_1/N_1 - n_2/N_2$. To construct a 95% confidence interval (CI) for θ, both exact and approximation methods are well studied (see, e.g., Chan and Zhang 1999; Chen 2002; Newcombe 1998). When the sample sizes for studying low incidence AEs are relatively large, the exact methods may not be feasible computationally. Among the approximation methods, Miettinen and Nurminen's (1985) method performs well as studied in Newcombe (1998) for proportion difference and Liu et al. (2006) for EAIR difference. Specifically, for proportion difference, the 95% CI is solved from

$$\chi^2_{0.95} = \frac{(\hat{p}_1 - \hat{p}_2 - \delta)^2}{\tilde{V}}$$

where $\hat{p}_i = n_i/N_i$, $\tilde{V} = \left[\frac{\tilde{p}_1(1-\tilde{p}_1)}{N_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{N_2}\right]\frac{N_1+N_2}{(N_1+N_2-1)}$, and $\tilde{p}_1$ is the maximum likelihood estimate of p_1 under $\tilde{p}_1 - \tilde{p}_2 = \delta$.

For the difference of two EAIRs, it is assumed that n_1 and n_2 follow independent Poisson distributions with parameters $\lambda_1 T_1$ and $\lambda_2 T_2$, respectively. The point estimate for the EAIR difference is $\hat{\theta} = n_1/T_1 - n_2/T_2$. Its 95% CI can be obtained similarly using the Miettinen and Nurminen's method (see Liu et al. 2006 for detailed formulas).

Example 1: A meta-analysis is conducted for a drug A (for confidentiality purpose the actual drug name is not disclosed) on comparing 3 gastrointestinal AEs that are clinically interested for the product. The meta-analysis included more than 17 trials, with duration of follow-up from 12 weeks up to 24 months. No recurrent events were collected in these studies. Table 1 summarizes the crude percentages of AEs and EAIRs, and their corresponding 95% CIs for the differences between drug A and control B.

Table 1 Meta-analysis of three gastrointestinal adverse events for drug (A) versus control (B)

AEs	Crude %			EAIR (per 100-patient year)		
	A N=3904 n (%)	B N=3310 n (%)	A-B % Diff (95% CI)	A n/T (rate)	B n/T (rate)	A-B Rate Diff (95% CI)
Nausea	67 (1.7)	51 (1.5)	0.2 (-0.4, 0.8)	67/3085 2.2	51/2480 2.1	0.1 (-0.7, 0.9)
Vomiting	47 (1.2)	38 (1.1)	0.0 (-0.5, 0.6)	47/3105 1.5	38/2496 1.5	0.0 (-0.7, 0.6)
Diarrhea	153 (3.9)	127 (3.8)	0.1 (-0.8, 1.0)	153/3019 5.1	127/2431 5.2	-0.2 (-1.4, 1.0)
N = number of patients, T = total follow-up time as defined in (2), n = number of patients with the AE.						

The crude percentages for the gastrointestinal AEs were low, around 1.1–3.9%. The percentages were similar between the test drug and control groups as the 95% CIs containing 0. However, because of different duration of follow-up in studies in the meta-analysis, it is not easy to interpret the crude percentage.

The EAIRs were around 1.5–5.2 per 100-patient year, which had a clear clinical interpretation. The values were slightly higher than the crude percentages because the averaged follow-up time for the studies in the analysis was less than a year. The EAIRs between treatment groups were also similar based on the 95% CIs.

Example 2: In a safety study for a vaccine, the primary endpoint is any serious adverse event (SAE) within 6 months post vaccination. In this study, the investigators were asked to make phone calls every 2 months to collect SAEs. Recurrent and multiple SAEs were collected throughout the study. In this analysis, we only consider the first SAE and summarize the crude percentage and EAIR in Table 2. The analysis of recurrent events will be considered in the next section.

With the deliberated follow-up plan in this vaccine study, most of the subjects completed the 6-month follow-up. Therefore, the overall follow-up time in patient-year was about half of the total number of subjects in the study. The values of EAIRs were about 2 times that of the crude percentages. The conclusions for the treatment comparisons were similar using either crude percentage or EAIR. No significant difference was seen from the analyses.

3.2 Comparing Overall Incidence Rates for Counts Data

When recurrent events are collected, both count of AEs and duration of follow-up are considered in statistical models for estimating the overall incidence rates. It is

Table 2 Analysis of serious adverse events for vaccine (A) versus control (B)

Crude %			EAIR (per 100-patient year)		
A N=3904 n (%)	B N=3310 n (%)	A-B % Diff (95% CI)	A n/T (rate)	B n/T (rate)	A-B Rate Diff (95% CI)
341 (5.7)	299 (5.0)	0.7 (-0.1, 1.5)	341/2880 (11.8)	299/2895 (10.3)	1.5 (-0.1, 3.1)
N = number of patients, T = total follow-up time as defined in (2), n = number of patients with the AE.					

likely that the recurrent events may be correlated; in that case the Poisson regression may not be appropriate because of over- or under-dispersion problem. Another issue for low incidence AEs is that majority patients might not report any AEs, which results in zero-inflated distribution.

To address the over-dispersion and zero inflation, several statistical models are proposed in the literature, including Poisson model with over-dispersion, negative binomial model, Poisson model with zero inflation, and negative binomial model with zero inflation (e.g., Bae et al. 2005). For these models, it is important to understand the meaning of the parameters in the models. Suppose μ is the mean for counts Y, then a Poisson regression is often represented by a log linear model

$$f(y) = \mu^y e^{-\mu}/y!, y \geq 0$$
$$\log(\mu) = \alpha + \beta I_{(trt=A)}, \text{var}(Y) = \mu$$

where $I_{(\text{E})} = 1$ if event E is true, 0 otherwise. From this model, the means for treatment groups A and B are $e^{\alpha+\beta}$ and e^{α}, respectively, and β is the log risk ratio between treatment groups A and B.

With similar notations, the other models can be written as follows.

1. Poisson model with over-dispersion (P w OD)

$$Y \sim \text{Poisson}(\mu), \text{var}(Y) = \varphi\mu$$

where $\varphi > 0$ is the dispersion parameter.

2. Poisson model with zero inflation (ZIP)

$$f(y) = \begin{cases} \omega + (1-\omega)\text{e}^{-\mu}, & y = 0 \\ (1-\omega)\mu^y \text{e}^{-\mu}/y!, & y > 0 \end{cases}, \quad \text{logit}(\omega) = \xi + \eta I_{(trt=A)}$$

where ω is a probability for excessive zero.

Table 3 Incidence, incidence ratio, and incidence difference under different statistical models

Model	Incidence for treatment A and B		Incidence Ratio	Incidence Difference
	μ_A	μ_B		
Poisson Poisson w OD NB	$e^{\alpha+\beta}$	e^{α}	e^{β}	$e^{\alpha+\beta}-e^{\alpha}$
ZIP ZINB	$(1-\text{logit}^{-1}(\xi+\eta))e^{\alpha+\beta}$	$(1-\text{logit}^{-1}(\xi))e^{\alpha}$	$\frac{\mu_A}{\mu_B}$	$\mu_A-\mu_B$

3. Negative binomial model (NB)

$$f(y)=\Gamma(y+1/k)(k\mu)^y/\Gamma(y+1)/\Gamma(1/k)/(1+k\mu)^{y+1/k},\quad y\geq 0$$

where k > 0 is a shape parameter for NB model.

4. Negative binomial model with zero inflation (ZINB)

$$f(y)=\begin{cases}\omega+(1-\omega)(1+k\mu)^{-1/k}, & y=0\\ (1-\omega)\Gamma(y+1/k)(k\mu)^y/\Gamma(y+1)/\Gamma(1/k)/(1+k\mu)^{y+1/k}, & y>0\end{cases}$$

In all these models, the meaning of the regression parameters α and β can be different. As summarized in Table 3, for Poisson, Poisson w OD, or NB models, β is the log of incidence ratio and $e^{\alpha+\beta}-e^{\alpha}$ is the incidence difference. For zero-inflated models (ZIP or ZINB), the incidence ratio or incidence difference is also related to the zero inflation parameters ξ and η. Therefore, we cannot simply test parameter β alone for comparing the treatment difference. A more appropriate comparison should be based on the incidence ratio or difference between groups, and constructing corresponding confidence intervals.

In order to calculate confidence interval for the incidence difference under these models, we consider the effective sample size approach (Li et al. 2006; Liu 2012). Suppose $\hat{\mu}$ and $SE(\hat{\mu})$ are point estimate and SE for μ of a given treatment group from an analysis model, we define effective follow-up and events as

$$\hat{T}=\hat{\mu}\ /\ [SE(\hat{\mu})]^2,\ \hat{n}=\hat{T}\times\hat{\mu}.$$

Apply this formula to both treatment groups A and B, we will get $(\hat{n}_A,\hat{T}_A)$ and $(\hat{n}_B,\hat{T}_B)$, then the 95% CI for incidence difference is obtained using Miettinen and Nurminen's method.

Remark: Among these models, it is suggested that the NB-based models may be more natural to account for over-dispersion (Keene et al. 2007). The NB model

assumes that the count for each individual follows a Poisson process with different incidence rate, which, as a set, are distributed according to a gamma distribution.

4 Application

In the vaccine study example described above, a subject may report multiple SAEs during the 6-month follow-up. Table 4 summarizes the number of subjects reporting SAEs by number of SAEs reported. It shows that quite a few subjects reported more than 1 SAEs.

To account for all SAEs observed in the study, a naïve estimate for the overall incidence rate would be the total SAEs divided by the total follow-up time. Its 95% CI can be obtained using the Miettinen and Nurminen's method (see Liu et al. 2006). We also use statistical models described in the previous section in this data. Table 5 presents the analysis results including the naïve method.

The results show that the naïve method is exactly the same as that from Poisson regression. Without considering the potential over-dispersion, the variability from the Poisson regression or zero-inflated Poisson model may be underestimated. As a result, the 95% CIs from those models were narrow and did not contain 0, which may lead to conclude that the rates were different between treatment groups. However, when over-dispersion was considered in the models of Poisson with OD, NB, or ZINB, the resulted 95% CIs all contained 0 implied no significant difference between the treatment groups.

Table 4 Number of subjects by reported number of SAEs and treatment group

# of SAEs	1	2	3	4	5	6	8	10	13
A	238	61	16	16	6	1	1	1	1
B	214	41	30	9	1	4	0	0	0

Table 5 Estimated overall incidence rates and incidence rate difference by statistical method

Method	Incidence Rates and Difference (per 100-person-year)		
	A	B	A-B (95% CI)
Naïve method	18.2	15.2	3.0 (0.9, 5.1)
Poisson	18.2	15.2	3.0 (0.9, 5.1)
ZI Poisson	18.5	15.4	3.1 (0.2, 6.0)
Poisson w OD	18.2	15.2	3.0 (-0.3, 6.4)
NB	18.6	15.6	3.1 (-.0.1, 6.3)
ZI NB	18.7	15.6	3.1 (-0.1, 6.3)

5 Conclusion Remarks

For low incidence AEs, a study with large sample size and long-term follow-up, or meta-analysis which combines multiple trials may be used. In this case, it is important to consider exposure and account for recurrent events in the analyses.

To account for exposure, EAIR is one commonly used measure for the incidence based on time-to-first event. To further account for recurrent events, an overall incidence rate may be estimated from statistical models such as Poisson or negative binomial regression. In many applications, a simple overall incidence rate or Poisson model may not be appropriate due to over-dispersion. A model adjusted for over-dispersion and zero inflation may be more suitable, e.g., zero-inflated Poisson model with over-dispersion, or zero-inflated negative binomial model. Some alternative models such as zero-inflated hurdle models may also be considered (see Yang et al. 2012).

It is noticed that the analysis based on exposure-adjusted incidence or overall incidence requires a constant hazard assumption, which may be approximately reasonable when the incidence rate is low. However, when the constant hazard is concerned in an application based on clinical or biological reasons, survival analysis methods such as Kaplan-Meier estimate or parametric time-to-event models may be considered. In this case, a simple summary statistic such as average incidence rate over the study period may not carry sufficient information to characterize the safety profile. The duration of follow-up, cumulative probability, or hazard function over time may be considered to provide a more complete profile of safety. Recently, a nonparametric approach called the mean cumulative function (MCF) has been proposed to provide statistical inference on recurrent AE profiles in randomized clinical trials. The details can be found in Siddiqui (2009) and Barker (2010). Finally, a related topic to low incidence AE analysis is safety signal detection, which is out of the scope for this paper. Many research papers are available in pharmacovigilance literature (see, e.g., Almenoff et al. 2005).

References

Almenoff J, Tonning JM, Gould AL, Szarfman A, Hauben M, Ouellet-Hellstrom R, Ball R, Hornbuckle K, Walsh L, Yee C, Sacks ST, Yuen N, Patadia V, Blum M, Johnston M, Gerrits C, Seifert H and LaCroix K (2005). Perspectives on the use of data mining in pharmacovigilance. Drug Safety, 2005: 981-1007.

Bae S, Famoye F, Wulu JT, Bartolucci AA and Singh KP (2005). A rich family of generalized poisson regression models with applications. Mathematics and Computers in Simulation 69: 4-11.

Barker C (2010). Exploratory method for summarizing concomitant medication data – the mean cumulative function. Pharmaceutical Statistics 9: 331-336.

Chan ISF and Zhang Z (1999). Test-based exact confidence intervals for the difference of two binomial proportions. Biometrics 55: 1202-1209.

Chen X (2002). A quasi-exact method for the confidence intervals of the difference of two independent binomial proportions in small sample cases. Statistics in Medicine 21: 943-956.

Chuang-Stein C (1998). Safety analysis in controlled clinical trials. Drug Information J. 32:1363S–1372S.

Keene ON, Jones MRK, Lane PW and Anderson J (2007). Analysis of exacerbation rates in asthma and chronic obstructive pulmonary disease: example from the TRISTAN study. Pharmaceut. Statist. 2007: 89–97.

Li XM, Mehrotra DV and Barnard J (2006). Analysis of incomplete longitudinal binary data using multiple imputation. Statistics in Medicine 25: 2107-24.

Liu G, Wang J, Liu K and Snavely DB (2006). Confidence Intervals for an Exposure Adjusted Incidence Rate Difference with Applications to Clinical Trials. Statistics in Medicine 25:1275-1286.

Liu G (2012). A note on effective sample size for constructing confidence intervals for the difference of two proportions. Pharmaceut. Statist. 11: 163–169.

Miettinen O and Nurminen M (1985). Comparative analysis of two rates. Statistics in Medicine 4:213-226.

Newcombe RG (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in Medicine 17: 857-872.

Siddiqui O (2009). Statistical Methods to Analyze Adverse Events Data of Randomized Clinical Trials. Journal of Biopharmaceutical Statistics 19:889-899.

Yang J, Li X and Liu G (2012). Analysis of zero-inflated count data from clinical trials with potential dropouts. Stat in Biopharm Research. 4: 273-283.

Statistical Power to Detect Cardiovascular Signals in First-in-Human Trials: Is It Really Small?

Ouhong Wang, Mike Hale, and Jing Huang

Abstract It is widely accepted that, due to the small size of first-in-human (FIH) trials, safety signals are difficult to detect. The chances of detecting early signals in cardiovascular safety, including heart rate, blood pressure, QT prolongation, etc., have long been considered to be remote. However, much of this belief is based on an analysis involving pair-wise comparisons of very small cohorts. When dose is considered as a continuous variable, dose–response becomes the main focus and power can be significantly improved with appropriate testing procedures. In this research, we try to quantify through simulations the power in this setting and demonstrate that cardiovascular safety signals in general have reasonable statistical power for early detection when using a dose–response analysis. The simulations account for different magnitudes of effects and various scenarios including linear, log-linear, and Emax relationships between dose and safety signal, together with multiple parametric and nonparametric tests.

1 Introduction

Cardiovascular safety signal is a very important aspect that we need to pay special attention to in all phases of clinical trials. Much effort has been applied to try detecting such signals as accurately as possible and as early as possible. Yet, it has long been thought that the chances of detecting these signals are low in first-in-human (FIH) studies due to the inherently small sample sizes of the study design. Such belief often is based on analysis using pair-wise comparison between the placebo and a particular dose group without considering the possibility of dose–response. As illustrated in Fig. 1 using variability estimates from empirical data, in

O. Wang (✉) • M. Hale • J. Huang
Global Biostatistical Sciences, Amgen Inc, Thousand Oaks, CA 91320, USA
e-mail: owang@amgen.com

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_23,

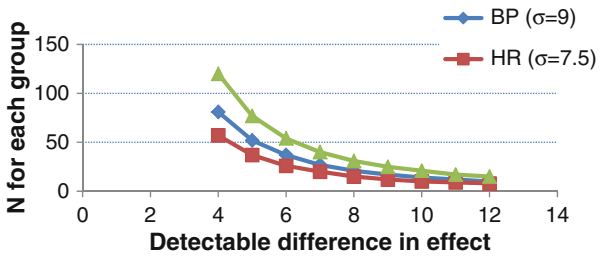


Fig. 1 Sample size required for 80% power at 2-sided $\alpha = 0.05$

order to achieve 80% power, pair-wise comparisons require sample sizes that are multiple-fold larger than those typically seen in FIH.

However, it is more informative to examine the safety signal relationship with dose as a continuum. When all doses in the FIH are considered in this manner, power naturally increases as the variability estimate for the dose–response relationship (e.g., slope in the linear dose–response case) is reduced compared to the intra-subject variability. It is not clear, though, what the magnitude of the power gain could be between these two analysis methods. In this research we use simulation to answer the question about how much better we could detect these important safety signals in the sense of power if we treat dose as a continuous variable and consider all doses together.

2 Method

2.1 Simulation Setup

The general framework of FIH studies is used. Specifically, we assume subjects are dosed sequentially in different cohorts, with each cohort using a dose higher than the previous one. Within each cohort, subjects are randomized to either placebo or active drug. Dose escalation is by a fixed factor, or equivalently, by a fixed amount on the log-scale.

The measurement of interest is change from baseline of a cardiovascular safety signal which could be heart rate, blood pressure, QT prolongation, etc. Assume that the change within each dose group follows a normal distribution and the variability (i.e., standard deviation) of these changes is SD=σ across all dose groups. Assume further that in the placebo group, the change from baseline has a mean of 0, and in the highest dose group, the change from baseline has a mean of A. We explore a range of A from the following: A = {0.2–2 by 0.2} $\times$ σ. In actual applications, data may be re-expressed, such as log transformation, if needed to yield approximate normality.

For a typical FIH study, doses of D, aD, a^2D, a^3D, …, a^{d-1}D are utilized in a cohort design with 3:1 active-to-placebo ratio. While this general setup allows flexibility in the dosing pattern, in this particular simulation we set a = 3, d = 6, and D = 1 which produce corresponding doses of 0 (placebo), 1, 3, 9, 27, 81, and 243. In this simulation, we examine cohort sizes of n = 4 and n = 8, which are two representative setups in FIH studies.

Nine models are considered to describe the true underlining relationship between the escalating dose and the change from baseline of each dose group: linear, log-linear, and E_{max} model following effect = dose × A/(ED_{50} + dose) where ED_{50} = 1/b of max dose, and b = {2, 3, 4, 5, 10, 20, 40}.

For each of these scenarios, the simulation is run for the two cohort sizes: n = 4 and n = 8; and the magnitude of mean change (for the highest dose group) is set at 10 evenly spaced points: 0.2σ, 0.4σ, 0.6σ, …., 2σ, with the magnitude of mean changes for the other doses dictated by the nine models above. After the mean effect for each dose is determined, we then simulate change from baseline values 1,000 times under normal distribution with the estimated mean and the common σ. Power, defined as the proportion of the 1,000 trials that reach statistical significance at 2-sided $\alpha = 0.05$, is calculated using three different statistical tests:

1. Jonckheere-Terpstra test (J-T).
2. Trend test for the slope of the linear dose–response being 0.
3. Trend test for the slope of the log-linear dose–response being 0.

While the latter two use linear regression models to test specific trends between dose and response, the J-T test is a nonparametric trend test. The alternative hypothesis for the J-T test is that there is *a priori* ordering of the populations. Letting θ_i be the population median for the ith population (i = 1 to K), the null and alternative hypotheses can be conveniently expressed as:

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_K$$

$$H_A : \theta_1 \leq \theta_2 \leq \ldots \leq \theta_K$$

with at least one inequality being strict. The test statistic derives from a comparison of how many data points are in order versus out of order in the sense of the alternative hypothesis. Approximation to standard normal distribution is used to infer test significance (Hollander et al. 1973; Kotz et al. 2006).

Note that the placebo subjects are combined to form one placebo group during analysis: for cohort size of n = 4, the placebo group will have 6 samples while the rest of the dose groups have 3 each; for cohort size of n = 8, the placebo group will have 12 samples while the rest of the dose groups have 6 each. Also note that for the log-linear trend test, we assign a hypothetical dose to the placebo group as D/a to make the log transformation feasible and the gap between doses evenly spaced in the log-scale. In this particular simulation D/a = 1/3. In addition, using these models, the power is independent of σ; and for the ease of implementation we take $\sigma = 7.5$, which is a reasonable value for heart rate variability in FIH studies observed.

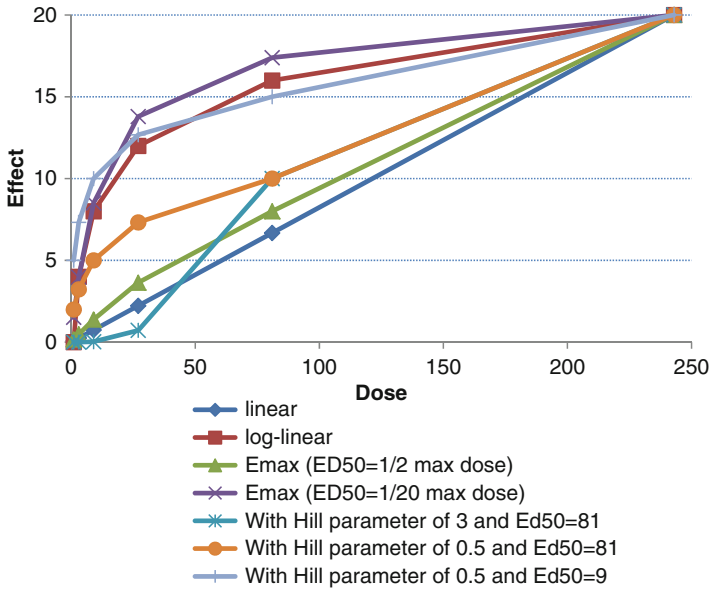


Fig. 2 Various dose-response scenarios

2.2 Adding Hill Parameter in the Emax Model

To investigate further the patterns between linear and log-linear trend we introduce the Hill parameter in the Emax model to provide a gently increasing S-shaped curve:

$$Effect = \frac{dose^H \times A}{ED_{50}^H + dose^H}$$

Figure 2 shows the effect of adding the Hill parameter. When ED_{50} is small, the dose–response behavior is similar to the log-linear trend. When ED_{50} is large and H is large, effect goes up slowly at the beginning and increases more rapidly in the middle dose range, exhibiting a more obvious S-shape which is overall better approximated by linear trend than anything else. When ED_{50} is large and H is small, the pattern is in-between linear and log-linear curves.

To understand the behavior of power in the presence of Hill parameter, we first looked at the Emax model with Hill parameter $h = 2$, sample size of 2:6 for each cohort, and effect of maximum dose being 1σ. We examined three ED_{50} cases where it falls on 9, 27, and 81, respectively. We then fixed the ED_{50} at a relatively large value of 81 and examined a range of Hill parameters spanning from 0.5 to 5 with 0.5 increments.

3 Results

The simulation results of the nine scenarios are summarized graphically that represent the power for all combinations of A, n, effect scenario, and statistical tests. Specifically, there will be one plot for each scenario, thus nine plots in total. Within each plot, the x-axis is the magnitude of the drug effect from 0.2σ to 2σ for the highest dose, and the y-axis is the estimated power from the simulation (i.e., the proportion of the 1,000 tests that has 2-sided p-value <0.05). We use solid lines to indicate the results from the smaller cohort size of $n = 4$ and dashed lines to indicate the results from the larger cohort size of $n = 8$. Different colors are used to indicate different tests employed: red for linear trend test; green for log-linear trend test, and blue for the Jonckheere-Terpstra test. Figure 3 represents the linear relationship; Fig. 4 represents the log-linear relationship; Figs. 5–11 represent the Emax model where $ED50 = 1/b$ of max dose, and $b = \{2, 3, 4, 5, 10, 20, 40\}$.

When the underlining dose–response relationship is linear, the linear trend test should naturally perform the best and the graph shows that clearly. With cohort size 4 (i.e., solid lines), when the effect is small (0.2σ), the power across the three tests is all low (around 5%). The difference becomes more noticeable with increasing effect size, and the advantage of the appropriate linear test becomes obvious with moderate increase in effect size, while the performance of log-linear test and J-T test remains similar until around 0.8σ. At the maximum effect of 2σ, the power of each test also reaches its highest, respectively, and the difference among tests is the largest as well: 85.2% for linear trend test, 58.9% for log-linear trend test, and

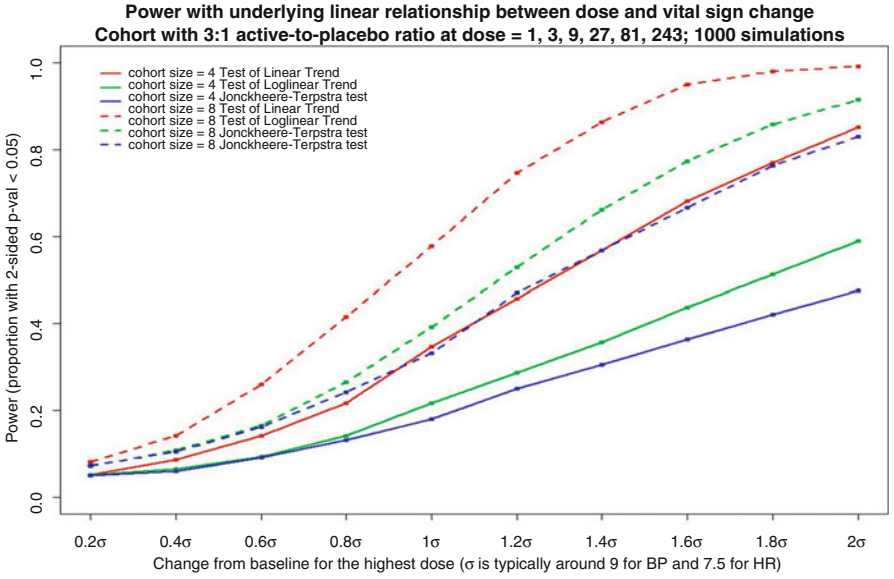


Fig. 3 Linear trend

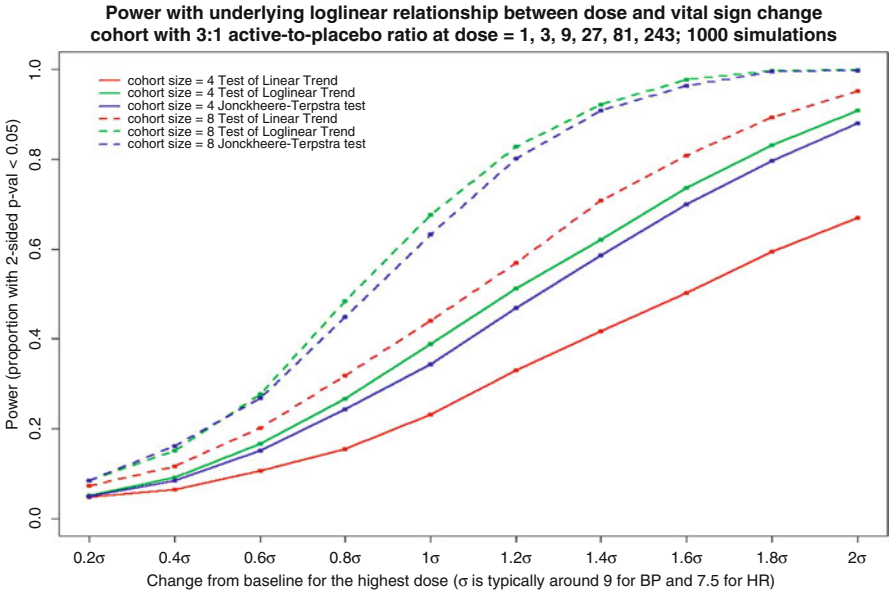


Fig. 4 Log-linear trend

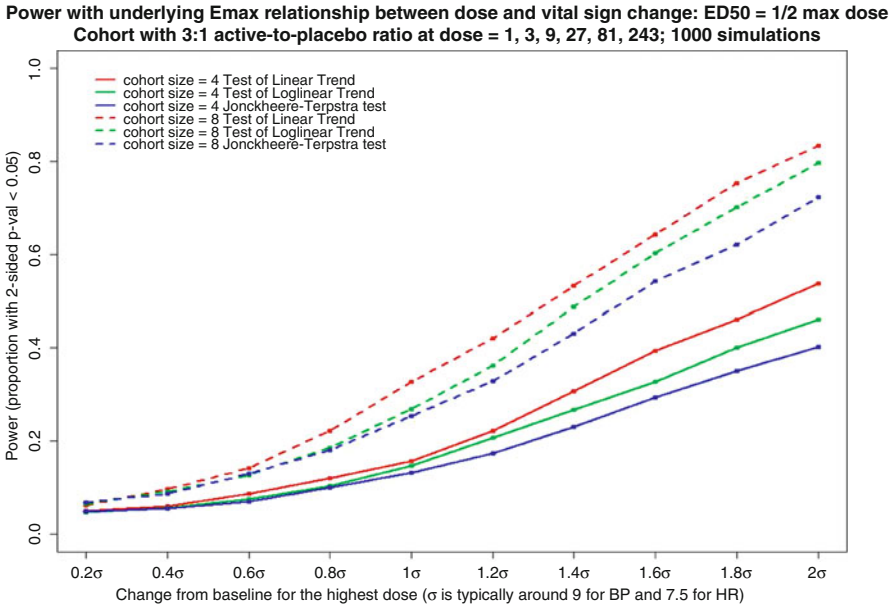


Fig. 5 Emax model with $b = 2$

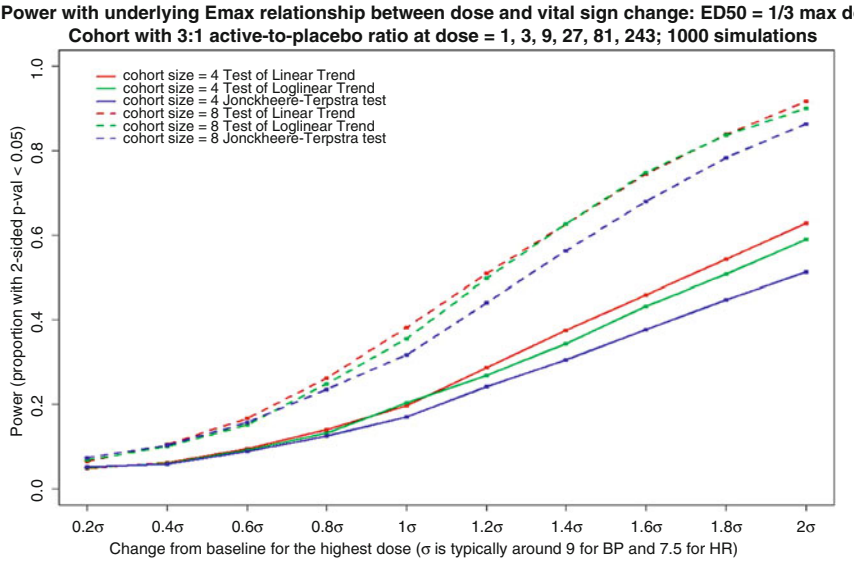


Fig. 6 Emax model with b = 3

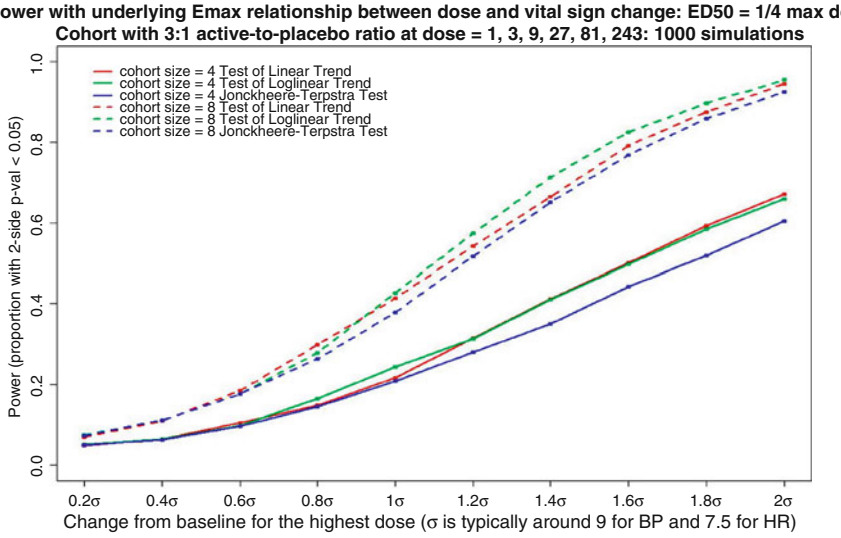


Fig. 7 Emax model with b = 4

47.6% for J-T test. When the sample size doubles from n = 4 to n = 8, the power increases as well. The linear test consistently provides the best power, while the log-linear test is slightly better than the J-T test. The biggest advantage of the linear test is around 1.2–1.4σ, after which, the power of the linear test gradually plateaus.

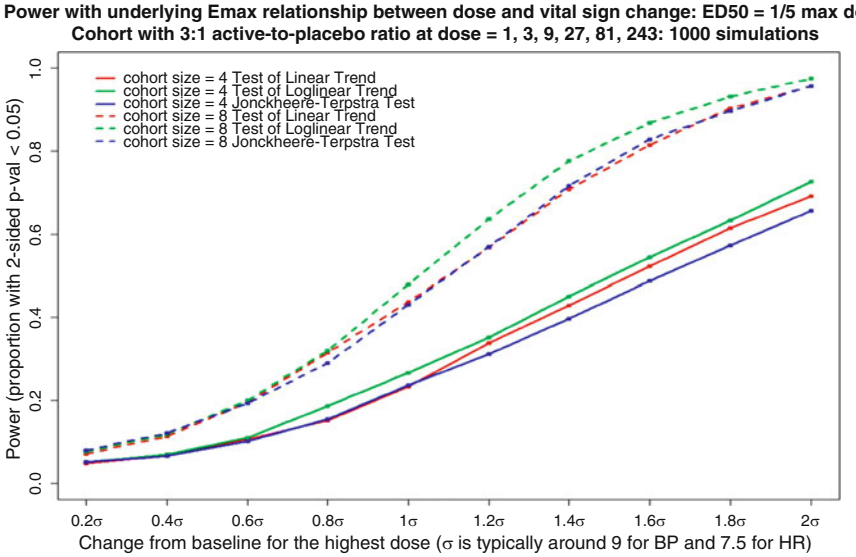


Fig. 8 Emax model with b = 5

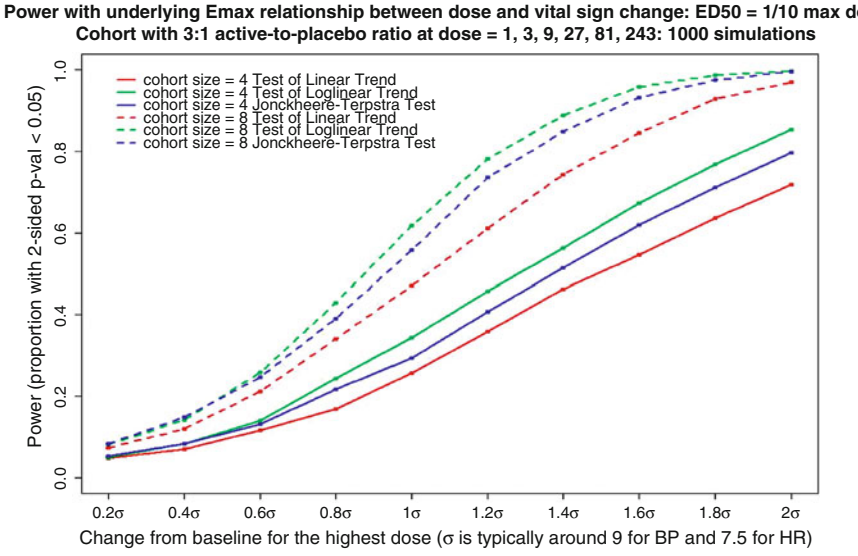


Fig. 9 Emax model with b = 10

Similar to the linear case, when the underlining dose–response relationship is log-linear, the log-linear trend test should naturally perform the best and the graph again shows that clearly. In this case, the log-linear test and the J-T test actually provide similar power with log-linear test being slightly and consistently better,

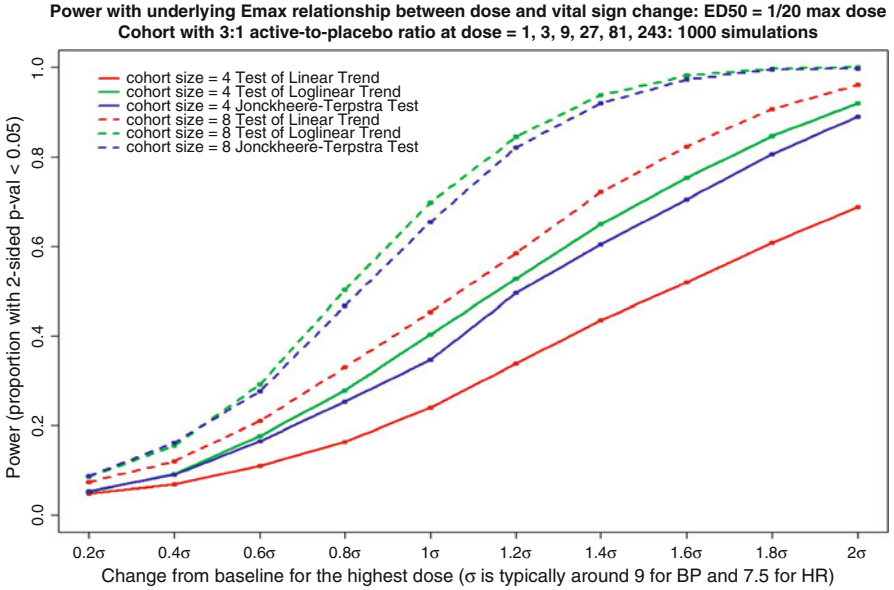


Fig. 10 Emax model with b = 20

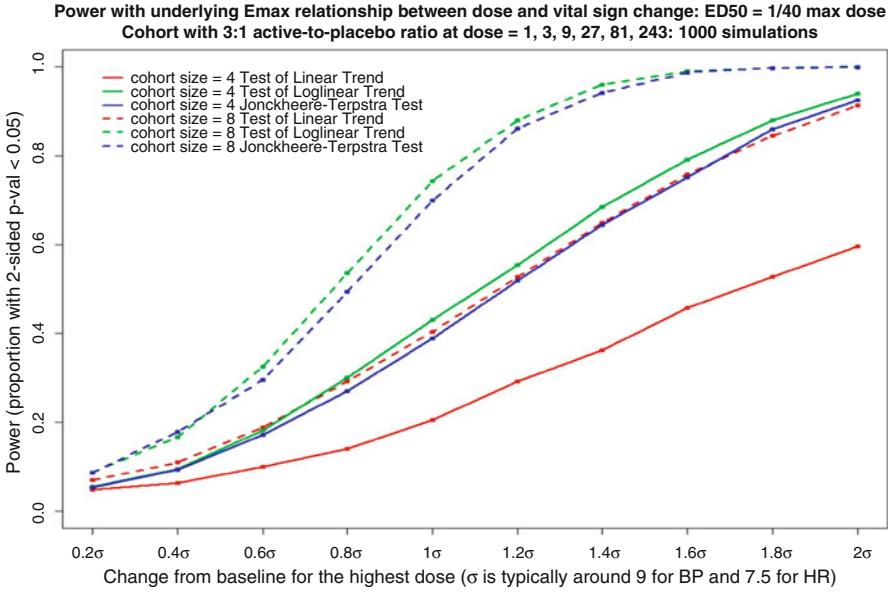


Fig. 11 Emax model with b = 40

while the linear test has substantially worse power. With the smaller cohort size, the difference between the log-linear test and the J-T test is less than 5% across the

whole range of effect sizes. The drop-off of power for the linear trend test could be as high as 23.9% compared to log-linear test and 21.2% compared to the J-T test. When the cohort size doubles to n = 8, the power increases as well. Now the difference between the log-linear trend test and the J-T test is further reduced with the biggest difference being 4.4% and in the majority of the cases the difference is less than 1.5%. The drop-off of power for the linear test is now 25.9% compared to the log-linear test and 23.1% compared to the J-T test. With the smaller cohort size, the log-linear test achieves power higher than 80% (power = 83.2%) at effect size 1.8σ. The linear trend test only reaches power of 66.9% with an effect size of 2σ. With double the cohort size, the log-linear test achieves power higher than 80% (power = 82.9%) at effect size of 1.2σ and power higher than 90% when the effect size is larger than 1.4σ. The J-T test reaches a power of 80.1% when the effect size is 1.2σ, and over 90% when the effect size is larger than 1.4σ. As a comparison, the linear trend test only reaches power over 80% with effect size of 1.6σ and 90% with the largest effect size of 2σ. In the set of scenarios using Emax model, effect goes up in an E_{max} fashion as following, where b = 2, 3, 4, 5, 10, 20, 40:

$$Effect = Dose \times \frac{A}{ED_{50} + Dose} = Dose \times \frac{A}{\frac{MaxDose}{b} + Dose}$$

When b is very small (b = 2), the dose–response behavior is more like the linear scenario. Not surprisingly, the linear test shows the best power followed by the log-linear test, then the J-T test. As b gradually increases to 3, 4, and 5, the power of the three tests is similar, while the sample size drives the difference in power. It is noticeable though that the linear trend test also gradually loses power relative to the other two tests. When b is 10, 20, and 40, the dose–response relationship is closer to the log-linear pattern and the log-linear test renders the best power, followed by the J-T test, while the linear test performs more and more poorly. The difference in power between the linear test and the other two is bigger as b gets larger. Note that the log-linear trend test consistently performs better than the J-T test, although the power difference is not dramatic.

We then look at Emax model with Hill parameters. We first look at h = 2, with cohort size of 8 (6:2). We explore three ED_{50} cases: 9, 27, and 81. The relationship between the change from baseline and dose is shown in Fig. 12. The left panel shows the dose using log-scale on the x-axis, while the right panel shows the dose using raw scale.

The powers of the three tests on these different scenarios are all summarized in Table 1.

Among the Emax models with h = 2, except for the case of high ED_{50} (81), log-linear test has the best power, while J-T test is the second best. For the case of ED_{50} dose = 81, linear test has the best power followed by log-linear test. In all five scenarios, J-T test never produces the best power. When h is less than 1, we use h = 0.5 as an illustrating example, the trend is very close to log-linear, while the magnitude of change (i.e., the slope and the change at the highest dose) is substantially smaller compared to Emax model with greater than 1 hill parameter

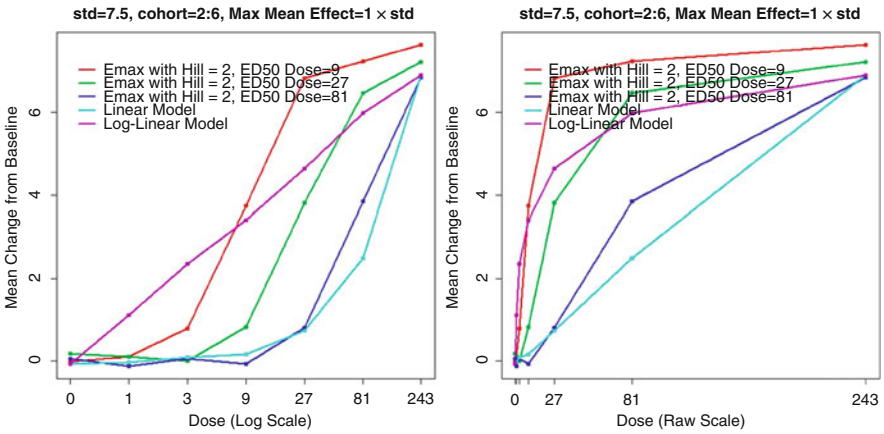


Fig. 12 Change from baseline vs. dose using log scale

Table 1 Power of three tests using different Hill parameters and different ED50

Power	Emax (h=0.5, ED_{50}=9)	Emax (h=0.5, ED_{50}=27)	Emax (h=0.5, ED_{50}=81)	Emax (h=2, ED_{50}=9)	Emax (h=2, ED_{50}=27)	Emax (h=2, ED_{50}=81)	Linear Model	Log-linear Model
Linear Test	0.297	0.246	0.239	0.509	0.583	0.551	0.527	0.4
Log-linear Test	0.528	0.386	0.31	0.815	0.655	0.41	0.345	0.621
J-T Test	0.49	0.376	0.289	0.775	0.601	0.36	0.296	0.57

and the same ED_{50}. As a result, the overall power is worse compared to the $h = 2$ scenarios. The log-linear test and the J-T test provide the best performance with the log-linear test being slightly and consistently better, while the linear test performance is substantially worse. The fact that the J-T test never produces the best power may seem surprising at the first look. Recall that our simulation setup had the low doses at 1, 3, 9, and 27. With a uniform standard deviation of σ and the maximum effect size of 2σ at the highest dose of 243, at the low end of the dose range there is considerable amount of overlap in the signal. Such scenarios render the J-T test less efficient in the low dose range and impact its overall performance. Varying model parameters to generate scenarios less favorable to the parametric models might be a good extension of this research to better understand J-T performance more broadly.

One observation is that when ED_{50} is large, log-linear test starts to lose its advantages. To investigate further, we fix ED_{50} at 81 and change the Hill parameter from 0.5 to 5. The result is summarized in Fig. 13. When $h = 0.5$, the trend is very similar to the log-linear trend. When h increases, the change from baseline increases much slower compared to log-linear trend before ED_{50} and much faster after ED_{50}. We calculate the power for the case of $h = 4$ as an example and the results are: linear test power = 62%, log-linear test power = 44%, and J-T test power = 37.1%. Overall, when h is small and/or ED_{50} is relatively low, the Emax model is very

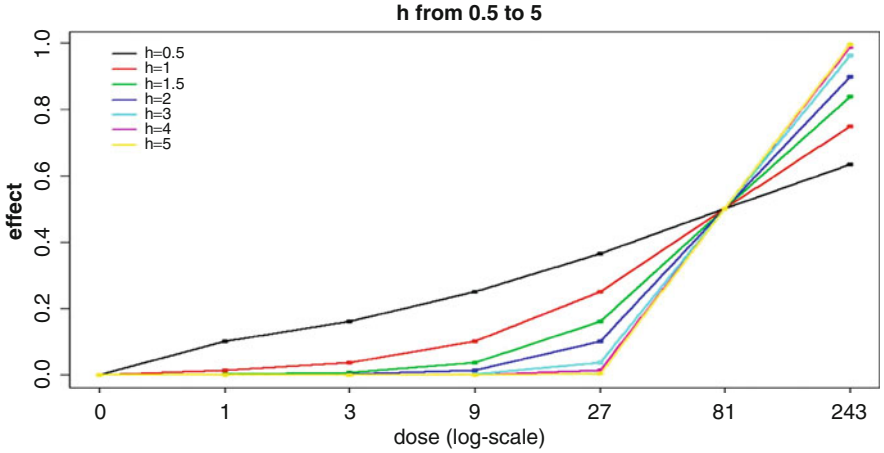


Fig. 13 Comparison of scenarios with different Hill parameters

similar to the log-linear model and thus the log-linear test performs the best. When ED_{50} is large and h is large as well, the Emax shape is very similar to the linear trend and thus the linear test does the best. When ED_{50} is large and h is small, the pattern falls in-between linear and log-linear.

4 Discussion

From this simulation, we see that in first-in-human studies, although pair-wise comparison for any dose vs. placebo has low power due to the limited sample size, taking all doses into consideration can increase the power substantially. It is important to have a good understanding of the relationship between the escalating dose and the change from baseline drug effect and use the appropriate test thereafter. Visualization between the dose and effect could help achieve such goal. When the relationship is unclear, log-linear dose trend test could be considered as the test of choice since in most cases, log-linear dose trend test consistently provides best or close-to-best power compared with other tests.

In addition, the following could be considered individually or in combination to further improve power. Multiple post-baseline measures can be combined to reduce variability. When PK profile is available, it can help select only relevant time points, again reducing variability. Baseline measures can be considered as covariates in the tests, eliminating the noise caused by them. When multiple small studies are available, they can be pooled, at least the placebo data, to increase sample size.

References

Hollander, M. and Wolfe, D. A.: Nonparametric Statistical Methods, pp.120-123. Wiley, New York (1973)

Kotz, S., Read, C. B., Blakrishnan, N., and Vidakovic, Brani. (eds): Encyclopedia of Statistical Sciences 2nd Edition, Volume 9, pp. 5811-14. Wiley, New York (2006)

Part VII
Longitudinal and Spatial Data Analysis

Constructing Conditional Reference Charts for Grip Strength Measured with Error

Pedro A. Torres, Daowen Zhang, and Huixia Judy Wang

Abstract Muscular strength, usually quantified through the grip strength, can be used in humans and animals as an indicator of neuromuscular function or to assess hand function in patients with trauma or congenital problems. Because grip strength cannot be accurately measured, several contaminated measurements are often taken on the same subject. A research interest in grip strength studies is estimating the conditional quantiles of the latent grip strength, which can be used to construct conditional grip strength charts. Current work in the literature often applies conventional quantile regression method using the subject-specific average of the repeated measurements as the response variable. We show that this approach suffers from model misspecification and often leads to biased estimates of the conditional quantiles of the latent grip strength. We propose a new semi-nonparametric estimation approach, which is able to account for measurement errors and allows the subject-specific random effects to follow a flexible distribution. We demonstrate through simulation studies that the proposed method leads to consistent and efficient estimates of the conditional quantiles of the latent response variable. The value of the proposed method is assessed by analyzing a grip strength data set on laboratory mice.

P.A. Torres
Department of Mathematical Sciences, University of Puerto Rico at Mayagüez, Mayagüez, Puerto Rico 00681-9018
e-mail: pa_torress@yahoo.es

D. Zhang • H.J. Wang (✉)
Department of Statistics, North Carolina State University, Campus Box 8203, Raleigh, NC, 27695-8203, USA
e-mail: dzhang2@stat.ncsu.edu; hwang3@ncsu.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_24,

1 Introduction

Grip strength is an indicator of neuromuscular function and is often used to assess the hand function in patients with trauma or congenital problems [19]. However, grip strength, particularly in mice and rats, is difficult to measure accurately [1]. As a consequence, researchers often collect between 2 and 5 measurements on the same subject because of the cost, effort, and time involved. These repeated measurements are contaminated with measurement error. A research interest in grip strength studies is estimating the conditional quantiles of the latent grip strength, say Y_i^*, which can be used to construct conditional grip strength charts. Those reference charts can help monitor the growth, follow progress after treatment, and identify abnormal cases. Reference charts for the latent grip strength are more broadly applicable than those based on the subject-specific averages of repeat measurements which are lab-specific, depending on the experiment protocols (e.g., repetition of measurements, dynamometers used, and muscle relaxation time). Our goal is to construct conditional reference charts for the latent grip strength given the contaminated replicates $\{Y_{ij}\}$ and covariates information. In other words, we are interested in estimating the conditional quantile function of the unobserved variable Y_i^*.

In the measurement error literature, researchers have discussed the estimation of quantiles of a latent variable when no covariates are present. In the context of density estimation, Caroll et al. [2] proposed to model the latent variable with a flexible distribution. Schechtman and Spiegelman [22] proposed two estimation methods, SIMEX [4] and efficient bootstrap combined with Jackknife. Li and Vuong [17] used the empirical characteristic function of pairs of repeated measurements to derive nonparametric estimators of the density of a latent variable. Delaigle et al. [7] suggested two deconvolution estimators for the density function of an unobserved variable based on contaminated replications. More recently, Tooze et al. [25] proposed a classical linear mixed effects model after applying a Box-Cox transformation to the contaminated replicates to estimate conditional quantiles of the unobserved response variable. None of the existing methods tackle the estimation of the conditional quantiles of Y_i^* using a quantile regression framework when the response has been measured repeatedly with error and covariates are present. Therefore, we propose a flexible semi-nonparametric method to estimate the conditional quantile function of a latent variable in a regression setting. We consider a two-stage model in which the regression errors for the latent variable follow a flexible distribution and the measurement errors are normally distributed.

2 Proposed Method

Let Y_i^* be a latent response variable and $\mathbf{x}_i$ a $p \times 1$ vector of covariates for the i^{th} subject, $i = 1, \ldots, m$. The vector $\mathbf{x}_i$ includes the constant 1 as the first element corresponding to the intercept term. For a given subject, the covariates in $\mathbf{x}_i$,

for instance age or weight, are assumed to remain constant across the repeated measurements. Our main objective is to estimate $Q_\tau(Y_i^*|\mathbf{x}_i) = \inf\{t \in \Re : P(Y_i^* \leq t|\mathbf{x}_i) \geq \tau\}$, where $0 < \tau < 1$ is the quantile level. We assume that Y_i^* and $\mathbf{x}_i$ are related through the following location-scale shift model

$$Y_i^* = \mathbf{x}_i^T \beta + (\mathbf{x}_i^T \gamma) u_i. \tag{1}$$

where β and γ are $p \times 1$ parameter vectors. In model (1) the covariates affect both the location and scale of the conditional distribution of Y_i^* given $\mathbf{x}_i$. To allow flexibility in model (1) we assume that the density of the regression errors u_i belongs to a class of smooth densities having a continuous and strictly increasing cumulative distribution function (CDF) denoted by $H_K(u_i, \mathbf{a})$, where $\mathbf{a}$ is the unknown parameter vector and K is a nonnegative integer that will be selected during the estimation process. A brief introduction to this class of smooth densities are deferred to the end of this subsection. The rationale behind this distributional assumption on u_i is that for some problems the unknown conditional distribution of the latent variable can be reasonably approximated by the members of this flexible semi-nonparametric (SNP) family. An advantage of this flexibility is that the bias due to misspecification could be diminished in many situations [23]. Another advantage of the proposed distribution is that the resulting likelihood function has a closed form and the computation is relatively straightforward.

The conditional linear quantile function resulting from model (1) is

$$Q_\tau(Y_i^*|\mathbf{x}_i) = \mathbf{x}_i^T \beta(\tau) = \mathbf{x}_i^T \{\beta + \gamma H_K^{-1}(\tau, \mathbf{a})\}, \tag{2}$$

provided that $\mathbf{x}_i^T \gamma > 0$ in the domain of $\mathbf{x}_i$. Although the location-scale shift model (1) assumes the conditional quantiles of Y_i^* are linear in $\mathbf{x}_i$, this model is convenient and has been widely used in quantile regression settings [11, 15, 21, 27]. Nonetheless, the proposed method can be extended to more general heteroscedastic models by replacing $\mathbf{x}_i^T \gamma$ with any parametric function $\sigma(\mathbf{x}_i, \gamma) > 0$.

If Y_i^* were observed, the quantile coefficients $\beta(\tau)$ in (2) can be estimated by the conventional quantile regression estimator [14, 15]. However, in many applications Y_i^* cannot be accurately measured. Instead, several repeated measurements $\{Y_{ij}\}_{j=1}^{n_i}$ are taken on the same subject i. For the observed response variable Y_{ij}, we assume the following classical measurement error model [2] $Y_{ij} = Y_i^* + e_{ij}$, where e_{ij} are i.i.d. $N(0, \sigma_e^2)$ random measurement errors that are independent of $\mathbf{x}_i$ and u_i. The model for the observed response according to (1) is then given by

$$Y_{ij} = Y_i^* + e_{ij} = \mathbf{x}_i^T \beta + (\mathbf{x}_i^T \gamma) u_i + e_{ij}. \tag{3}$$

Model (3) differs from the classical linear mixed model since the former includes a single term for the random effects u_i that could follow non-normal distributions, and its design matrix for the random effects includes the unknown parameters γ.

Some researchers apply conventional linear quantile regression (QR) method using the subject-specific average of repeated measurements, namely, $\bar{Y}_i =$

$n_i^{-1}\sum_{j=1}^{n_i} Y_{ij}$, as the response variable to estimate $Q_\tau(Y_i^*|\mathbf{x}_i)$ [18, 19, 26]. However, this naive QR approach suffers from model misspecification and often leads to biased estimates of the conditional quantiles of the latent response as well as the covariates effects.

In the proposed method we assume that the subject-specific random effects u_i follow a standard SNP distribution with density belonging to a class of smooth densities, introduced originally in the econometrics context [5, 9]. A univariate random variable u follows a standard SNP distribution if its density is given by $h_K(u,\mathbf{a}) = P_K^2(u,\mathbf{a})\phi(u) = (\mathbf{a}^T\mathbf{u})^2\phi(u)$, where $\mathbf{u} = (1,u,\cdots,u^K)^T$, K is the degree of the polynomial $P_K(\cdot)$, $\mathbf{a} = (a_0,a_1,\cdots,a_K)^T$ comprises the $(K+1)$ polynomial coefficients and $\phi(u)$ is the standard normal density. A comprehensive discussion of the SNP estimator and its properties can be found in [6, 9, 28].

2.1 Computational Details

Let $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{in_i})^T$ be the $n_i \times 1$ vector of repeated measurements for the i^{th} subject. The location-scale shift model (3) can be rewritten as

$$\mathbf{Y}_i = \mathbf{X}_i\beta + (\mathbf{X}_i\gamma)u_i + \mathbf{e}_i, \tag{4}$$

where $\mathbf{X}_i = \mathbf{1}_{n_i} \otimes \mathbf{x}_i^T$ is a $n_i \times p$ design matrix, $\mathbf{1}_{n_i}$ is a column vector of ones and $\otimes$ denotes the usual Kronecker product, u_i is a univariate random variable following a standard SNP distribution, and $\mathbf{e}_i$ is a $n_i \times 1$ vector that comprises the measurement errors associated with the i^{th} subject. In our implementation we assume that $\mathbf{e}_i \sim N(\mathbf{0},\sigma_e^2\mathbf{I}_{n_i})$, where $\mathbf{I}_{n_i}$ is the identity matrix of order n_i, but the proposed method can be extended to models with other measurement error distributions. The parameter vector of dimension $d = 2p+K+1$ is $\theta = (\beta^T,\gamma^T,\varphi^T,\sigma_e^2)^T$, where the K-dimensional vector φ results from the reparameterization of the polynomial coefficients $\mathbf{a}$. This reparameterization allows one to use unconstrained optimization techniques to compute the MLE of θ (see details in [3, 28]).

The log-likelihood function based on model (4) can be written as

$$l(\theta;\mathbf{Y}) = \sum_{i=1}^{m}\Big[\log g(\mathbf{Y_i};\theta) + \log E_{W_i|\mathbf{Y}_i;\theta}\big\{P_K^2(W_i)\big\}\Big], \tag{5}$$

where $g(\mathbf{Y}_i;\theta)$ is the normal density with parameters $E(\mathbf{Y}_i;\theta) = \mathbf{X}_i\beta$ and $Var(\mathbf{Y}_i;\theta) = \sigma_e^2\mathbf{I}_{n_i} + (\mathbf{x}_i^T\gamma)^2\mathbf{J}_{n_i}$, where $\mathbf{J}_{n_i}$ is the square matrix of ones of dimension n_i. Given $\mathbf{Y}_i$, the random variable W_i in (5) has a normal density with parameters $E(W_i|\mathbf{Y}_i;\theta) = \{n_i(\mathbf{x}_i^T\gamma)(\bar{Y}_i - \mathbf{x}_i^T\beta)\}/\{\sigma_e^2 + n_i(\mathbf{x}_i^T\gamma)^2\}$ and $Var(W_i|\mathbf{Y}_i;\theta) = \sigma_e^2/\{\sigma_e^2 + n_i(\mathbf{x}_i^T\gamma)^2\}$.

The log-likelihood (5) involves a normal-based likelihood and higher moments of a normal distribution that have closed forms. Therefore, the function (5) can be maximized using standard optimization algorithms such as Quasi-Newton Rapshon. Starting values for γ can be obtained by using $\bar{Y}_i$ and the median regression method

in [11]. For the angles φ, a grid of starting values on $(-\pi/2, \pi/2]^K$ usually deliver good results [28]. Otherwise, starting values can be easily obtained using a classical linear mixed model. The linearity of the conditional quantile (2) is guaranteed by the condition $\mathbf{x}_i^T \gamma > 0$. In practice, we found it unnecessary to enforce that constraint in the optimization routine if the location-scale shift model provides a reasonable fit to the data. Moreover, negative values of $\mathbf{x}_i^T \gamma$ resulting from the unconstrained optimization may be an indication of model misspecification [11].

The value of K in $h_K(u, \mathbf{a})$ can be selected by using adaptive rules such as BIC, AIC, or Hannan-Quinn (HQ) criterion [5, 10]. In our simulation studies we used the HQ criterion defined as $-l(\hat{\theta}; \mathbf{Y}) + d \log(\log N)$, where d is the dimension of θ and $N = \sum_{i=1}^{m} n_i$. In many applications values of $K = 0, 1, 2$ may be enough to approximate the true distribution of interest [3, 5, 28]. For highly skewed distributions, though, larger values of K are needed in order to get better approximations [8, 24].

2.2 Asymptotic Properties and Inference

The conditional quantile function of Y_i^* can be estimated by $\hat{Q}_\tau(Y_i^* | \mathbf{x}_i) = \mathbf{x}_i^T \hat{\beta}(\tau)$, where $\hat{\beta}(\tau) = \hat{\beta} + \hat{\gamma} H_K^{-1}(\tau; \hat{\varphi})$. For a given τ and $\hat{\varphi}$, the quantity $H_K^{-1}(\tau; \hat{\varphi})$ can be found by solving the nonlinear equation $H_K(u; \hat{\varphi}) = \tau$ for u. In our implementation we used the bisection method to compute the quantiles of the SNP distribution. The asymptotic properties of the proposed ML-SNP estimator $\hat{\beta}(\tau)$ are stated in the following theorem.

Theorem 1. *Let $\beta_0(\tau) = \beta_0 + \gamma_0 \xi_K(\tau; \varphi_0)$, $0 < \tau < 1$, be the true quantile coefficients in (2) for a known K, where $\xi_K(\tau; \varphi_0) = H_K^{-1}(\tau, \varphi_0)$. Then, under classical regularity conditions for the MLE of θ and the SNP density $h_K(\tau, \varphi)$, the estimator $\hat{\beta}(\tau) = \hat{\beta} + \hat{\gamma} \xi_K(\tau; \hat{\varphi})$ satisfies that: i) $\hat{\beta}(\tau) \xrightarrow{p} \beta_0(\tau)$; ii) $m^{\frac{1}{2}} \left\{ \hat{\beta}(\tau) - \beta_0(\tau) \right\} \xrightarrow{d} N(\mathbf{0}, \Omega_0)$, as $m \to \infty$, where Ω_0 is the asymptotic covariance matrix. If $K = 0$, then $\Omega_0 = \mathbf{D} \Sigma_{\theta_0} \mathbf{D}^T$, where $\mathbf{D} = \left[\mathbf{I}_p, \Phi^{-1}(\tau) \mathbf{I}_p, \mathbf{0}_p \right]$, $\mathbf{I}_p$ is the identity matrix of order p, $\mathbf{0}_p$ is a p-dimensional vector of zeroes, $\Phi(\cdot)$ is the CDF of a standard normal distribution, and Σ_{θ_0} is the asymptotic covariance matrix resulting from (5). If $K > 0$, then $\Omega_0 = g'(\theta_0) \Sigma_{\theta_0} g'(\theta_0)^T$, where $g'(\theta) = \left[\mathbf{I}_p, \xi_K(\tau, \varphi) \mathbf{I}_p, \gamma \partial \xi_K(\tau, \varphi) / \partial \varphi^T, \mathbf{0}_p \right]$ is the Jacobian matrix, and $g(\hat{\theta}) = \hat{\beta}(\tau)$.*

In Theorem 1, $\partial \xi_K(\tau, \varphi) / \partial \varphi^T = -\{2 / h_K(\xi, \mathbf{a})\} \mathbf{a}^T \Lambda \mathbf{B}^{-1} \mathbf{C}$, where Λ is a symmetric square matrix of order $K+1$, $\mathbf{C} = \partial \mathbf{c} / \partial \varphi^T$ of dimension $(K+1) \times K$, and $\mathbf{B}$ and $\mathbf{c}$ are defined in [28, p. 797]. The $(i,j)^{th}$ element of Λ, denoted by λ_{ij}, is defined as $\lambda_{ij} = \int_{-\infty}^{\xi} t^{i+j-2} \phi(t)\, dt$, $i, j = 1, 2, \ldots, K+1$, where $\phi(\cdot)$ is the density function of a standard normal distribution.

The matrix $g'(\theta_0)$ involves the SNP density $h_K(u, \mathbf{a})$, which may take zero values depending on its parameters. For instance, when $K = 1$ the density $h_K(u, \mathbf{a})$

takes a zero value when $u = -a_0/a_1$, $a_1 \neq 0$. Hence the formula based on Taylor expansion is unstable around the quantile levels at which the densities are close to zero. Moreover, according to our simulation studies, the formula based on Taylor expansion often underestimates the standard errors of $\hat{\beta}(\tau)$ for small sample sizes. We adopt a resampling method by perturbing the log-likelihood as an alternative to estimate the standard errors [12, 13, 20].

3 Simulation Studies

We assessed the performance of the proposed method through the model $Y_{ij} = Y_i^* + e_{ij}$, where $Y_i^* = (4+2x_i)+(1+0.5x_i)u_i$, $x_i \sim U(0,1)$, $e_{ij} \sim N(0,\sigma_e^2)$, and $u_i \sim N(0,1)$ or $u_i \sim 0.75N(-1,1)+0.25N(2,1)$. We considered the following factors (levels): number of subjects ($m = 100, 300$), number of repeated measurements ($n = 2, 4$), and measurement error ($\sigma_e = 0.5, 1, 2$). The values for σ_e lead to conditional reliability ratios for individual measurements ranging from 0.20 to 0.96. The conditional reliability ratio is defined as $\rho(x_i) = Var(Y_i^*|x_i)/Var(Y_{ij}|x_i)$. For a given τ, the true quantile coefficients are given by $\beta_0(\tau) = 4 + F^{-1}(\tau)$ and $\beta_1(\tau) = 2 + 0.5F^{-1}(\tau)$, where $F^{-1}(\cdot)$ is the inverse CDF of u_i.

We considered five estimators: *ML-SNP*, *ML-Normal*, *REML-Normal*, *omniscient*, and *naive*. The *ML-SNP* estimator assumes an SNP distribution for the subject-specific random effects u_i, while the *ML-Normal* and *REML-Normal* assume normality. The method *omniscient* is the conventional quantile regression method using the latent response Y_i^*. This estimator is infeasible in practice but it serves as a benchmark for comparison. The *naive* estimator results from the conventional linear quantile regression with $\bar{Y}_i$ as the response variable. We fitted the models using an SNP distribution with K chosen among $0, 1$, or 2, using the HQ criterion. These grid points of K cover a broad class of distributions while having the method computationally tractable. We obtained estimates of the quantile coefficients $\beta_0(\tau)$ and $\beta_1(\tau)$ for $\tau = 0.1, \ldots, 0.9$, using the five methods described above. We compared these methods using the average bias and mean squared error (MSE) ratio of the estimators based on 1,000 Monte Carlo replicates. The MSE ratio is defined as the MSE of a given estimator divided by that of the *omniscient* estimator. Finally, we evaluated the adequacy of the variance formula for the ML-SNP estimator introduced in Sect. 2.2.

Due to space limit, we only report results from two simulation scenarios. The results for other scenarios show similar patterns. Figure 1 shows the average bias and MSE ratio of the estimators for the normal case with a balanced design with $n = 4$, $m = 100$, and $\sigma_e = 1$. For normally distributed random effects, the HQ criterion chose $K = 0$ in about 83% and 95% of the times when $m = 100$ and 300, respectively. From Fig. 1, the SNP- and normal-based estimators are approximately unbiased. Naive method underestimates the intercept, while it overestimates the slope when $\tau < 0.5$. The bias is larger for quantiles on the tails of the distribution. Although the relative bias of naive estimates is small, it does not vanish as m increases. Moreover,

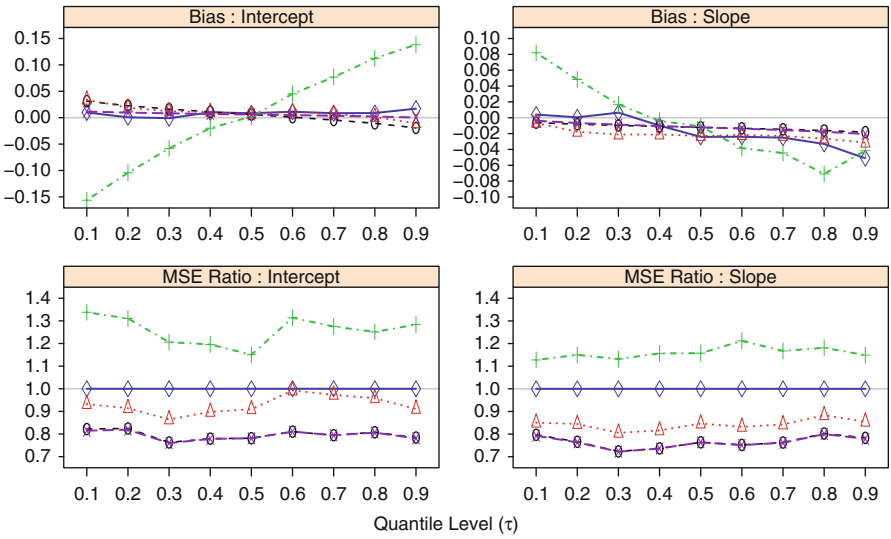


Fig. 1 Average bias and MSE ratio of quantile coefficient estimators with $u_i \sim N(0,1), m = 100, n = 4$. Omniscient (◇), ML-Normal (○), ML-SNP (△), Naive (+), REML-Normal (×). Reliability ratios: $\rho(0) = 0.50$ and $\rho(1) = 0.69$.

naive estimator has a systematic bias that increases with σ_e, regardless of the m value. Finally, all the methods yield unbiased estimates when $\tau = 0.5$. Note that the naive yields unbiased estimates at $\tau = 0.5$ because the conditional median of Y_i^* and $\bar{Y}_i$ are the same so the naive estimator does not suffer from model misspecification. In terms of the MSE ratio, the SNP estimators for random normal effects lose some efficiency when compared to the normal-based estimators, but that difference in MSE becomes negligible when the sample size increases.

For the normal mixture distribution, the HQ criterion chose $K = 0, 1$, and 2 in about 16 %, 22 %, and 62 % of the times when $m = 100$, whereas $K = 2$ was chosen 99.7 % of the times when $m = 300$. For the sake of clarity, we excluded the results of the normal-based estimators from Fig. 2 because of the considerable bias resulting from the misspecification of the subject-specific random effects distribution. From Fig. 2, the naive method underestimates the intercept for $\tau \leq 0.38$ and overestimates it when $\tau > 0.38$, which approximately corresponds to the quantile level where the distribution takes its mode. An opposite bias pattern occurs with the slope. The naive estimator has negligible bias at $\tau \approx 0.38$ since $\bar{Y}_i$ and Y_i^* have the same conditional quantiles at this quantile level. The ML-SNP and the omniscient estimator are approximately unbiased. The ML-SNP estimator shows smaller MSE compared to the other estimators. In particular, the ML-SNP estimator outperforms the omniscient estimator, since the latter is obtained nonparametrically without modeling the response distribution while the former estimates the normal mixture distribution satisfactorily. This fact reveals the advantage of the flexibility of the ML-SNP estimator to approximate distributions not belonging to the SNP family.

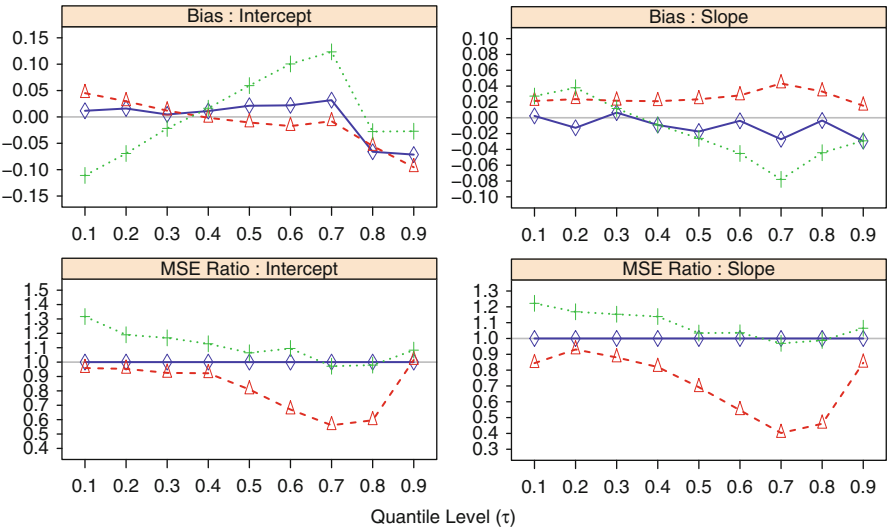


Fig. 2 Average bias and MSE ratio of quantile coefficient estimators with $u_i \sim 0.75N(-1,1) + 0.25N(2,1), m = 100, n = 4$. Omniscient (◇), ML-SNP (△), Naive (+). Reliability ratios: $\rho(0) = 0.73$ and $\rho(1) = 0.86$.

Table 1 Standard errors and coverage probabilities for the ML-SNP estimator of quantile coefficients when $m = 100$. *T*: variance estimation method based on Taylor expansion; *R*: resampling method; *Bias*: average bias; *MC SD*: Monte Carlo standard deviation; *SE*: average standard error; *CP*: coverage probability of 95 % Wald-type confidence interval.

		$\beta_0(\tau)$				$\beta_1(\tau)$			
$u_i \sim N(0,1)$									
τ	Method	Bias	SE	MC SD	CP	Bias	Ave. SE	MC SD	CP
0.1	T	0.034	0.346	0.367	0.919	-0.007	0.645	0.678	0.931
	R		0.339		0.940		0.624		0.940
0.5	T	0.007	0.245	0.256	0.930	-0.023	0.449	0.474	0.927
	R		0.252		0.935		0.461		0.935
$u_i \sim 0.75N(-1,1) + 0.25N(2,1)$									
0.1	T	0.045	0.365	0.400	0.910	0.021	0.688	0.731	0.915
	R		0.389		0.920		0.743		0.930
0.5	T	-0.011	0.319	0.395	0.866	0.023	0.565	0.720	0.857
	R		0.402		0.915		0.706		0.945
0.9	T	-0.095	0.532	0.741	0.826	0.015	0.957	1.343	0.831
	R		0.659		0.930		1.247		0.935

Table 1 summarizes the standard error estimates for some quantile levels based on the variance formula given in Sect. 2.2. Because the formula described the sampling variation of the estimator adequately when $m = 300$, we show only the results when $m = 100$. For small sample sizes, the proposed formula based on Taylor

expansion underestimates the true sampling variation of the ML-SNP estimator. The underestimation is more severe for the normal mixture case, particularly for upper quantiles. A possible explanation for the underestimation is that the random effects distribution may be misspecified when the sample size is small. This leads to poor estimates of the density value at the quantile of interest, and therefore, unsatisfactory performance of the proposed formula. Table 1 also shows the results using the resampling approach as an alternative to the Taylor expansion formula when $m = 100$. For each of the 200 simulated data sets, we generated 50 resamples based on the random perturbations generated from an exponential distribution with mean and variance equal to 1. The resampling approach gives better estimates of the standard errors, particularly for non-normal random effects and small sample sizes where the formula based on Taylor expansion does not perform adequately.

4 Application: Conditional Grip Strength Charts

We illustrate the proposed method using the mice data from the JaxWest1 study [16]. The study assessed the physiology of seven inbred mice strains. We focus on the relationship between grip strength and weight for illustration. Five measurements ($n = 5$) of the forelimb grip strength (g) were collected from $m = 112$ mice aged about 12 weeks so the total number of observations is $N = 560$. Our objective is to estimate the conditional quantiles of the latent grip strength as a function of weight. We propose the model $Y_{ij} = Y_i^* + e_{ij} = \beta_0 + \beta_1 \tilde{W}_i + (\gamma_0 + \gamma_1 \tilde{W}_i)u_i + e_{ij}$, where Y_{ij} and Y_i^* are the log-transformed observed and the latent grip strength, respectively. The $\tilde{W}_i$ is the weight centered at zero making the interpretation of the intercept quantile coefficient easier. The log transformation was adopted to satisfy the classical measurement error assumptions, which were verified using different tools. The values of the HQ criterion for $K = 0, 1$, and 2 were 0.066, 0.068, and 0.070, respectively. The SNP($K = 1$) is similar to the normal distribution, whereas SNP($K = 2$) is slightly left-skewed. A visual inspection to the quantile lines from the naive method in the log scale (Fig. 3) suggests that the lower quantiles are farther apart from each other than the upper quantiles, which indicates a skewed distribution. Therefore, we argue that an SNP($K = 2$) is more appropriate for this data set.

The estimates and their respective standard errors for the ML-SNP($K = 2$) are $\hat{\theta} = (\hat{\beta}^T, \hat{\gamma}^T, \hat{\varphi}^T, \hat{\sigma}_e^2)^T = (4.623, 0.025, 0.147, 0.011, 0.878, 0.450, 0.048)^T$ and $\widehat{SE}(\hat{\theta}) = (0.140, 0.008, 0.083, 0.004, 0.133, 0.309, 0.003)^T$. The SNP distribution has mean and variance equal to 0.703 and 1.508. The estimated reliability ratio is, for instance, $\hat{\rho}(25) = 0.39$ when the weight is 25g. The reliability ratio increases with weight because the variance of the latent grip strength is larger for heavier mice while the variance of the measurement error remains constant according to the proposed model. Figure 3 plots the fitted conditional quantile lines at $\tau = 0.05, 0.25, 0.50, 0.75, 0.95$ in the log scale using the naive and ML-SNP($K = 2$)

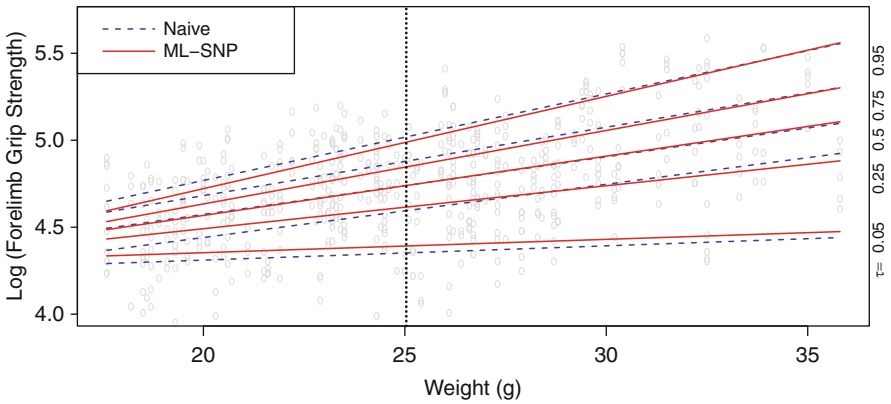


Fig. 3 Fitted conditional quantile lines in log scale for mice grip strength data. The circles represent the grip strength measurements. The vertical dashed line is the average weight.

approach. We can see that the naive method gives larger estimates of the upper quantiles of the latent grip strength, particularly for lighter mice. For instance, the naive estimate of the 75^{th} percentile is about the same as the ML-SNP estimate of the 95^{th} percentile when weight is about 18g. Conversely, the naive estimates are somewhat smaller for lower quantiles. The median grip strength line is similar in both methods. In this application the point estimates of the quantile coefficients using the naive and the ML-SNP($K = 2$) do not show significant differences mainly because of the considerable number of repeated measurements. In other studies, though, taking such a number of repeated measurements may be restrictive, and therefore, the naive estimates may exhibit larger bias in those situations.

5 Conclusions and Discussion

We proposed a flexible semi-nonparametric method to estimate the conditional quantile function of a latent response variable when it is measured repeatedly with error and covariates are present. The proposed method is flexible since the subject-specific random effects are assumed to follow a flexible SNP distribution. The method also accommodates heteroscedastic regression errors and is relatively easy to implement given the closed form of the likelihood function. For the location-scale shift model we derived the asymptotic properties of the proposed estimator. The proposed variance formula works well for moderate sample sizes. For small sample sizes, we recommend the proposed resampling method. Through the analysis of a lab mice data, the proposed method showed to be useful to estimate the conditional quantiles of the latent forelimb grip strength as a function of mice weight.

Simulation studies also suggest that our method outperforms the naive method for normal and non-normal cases across different measurement error scenarios.

Even for normal mixture distributions that do not belong to the SNP family, the proposed method has ignorable bias since the SNP distribution is flexible enough to provide decent approximations even with a small K. The ML-Normal and REML-Normal estimators are inappropriate when the distribution of the subject-specific random effects deviates considerably from normality, such as the normal mixture case considered in the simulation study.

One limitation of the proposed method is that for regression errors with heavy-tailed distributions, the SNP fit requires a larger polynomial degree K and this could be computationally troublesome. In such cases, one can apply a simple monotone transformation to the data to lighten the tails and fit an SNP with a smaller K. The conditional quantile in the original scale can be estimated directly by using its property of equivariance to monotone transformations.

Acknowledgements The research of Torres and Wang was supported by NSF award DMS-1007420 and NSF CAREER award DMS-1149355 and the research of Zhang was supported by the HIH grant R01 CA85848-12 and the NIH/NIAID grant R37 AI031789-20.

References

[1] Barreiro, E., Marin-Corral, J., Sanchez, F., Mielgo, V., Alvarez, F.J., Galdiz, J.B., Gea, J.: Reference values of respiratory and peripheral muscle function in rats. J. Anim. Physiol. Anim. Nutr. **94**, 393–401 (2010)
[2] Caroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M.: Measurement Error in Nonlinear Models: A Modern Perspective. Chapman & Hall/CRC, Boca Raton, FL (2006)
[3] Chen, J., Zhang, D., Davidian, M.: A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution, Biostatistics **3**, 347–360 (2002)
[4] Cook, J.R., Stefanski, L.A., Simulation-extrapolation estimation in parametric measurement error models, Journal of the American Statistical Association **89**, 1314–1328 (1994)
[5] Davidian, M., Gallant, A.R.: The nonlinear mixed effects model with a smooth random effects density, Biometrika. **80**, 475–488 (1993)
[6] Davidian, M., Giltinan, D.M.: Nonlinear Models for Repeated Measurement Data. Chapman & Hall, London; New York (1995)
[7] Delaigle, A., Hall, P., Meister, A.: On deconvolution with repeated measurements. The Annals of Statistics **36**, 665–685 (2008)
[8] Fenton, V., Gallant, R.: Qualitative and asymptotic performance of SNP density estimators. J. Econ. **74**, 77–118 (1996)
[9] Gallant, A.R., Nychka, D.W.: Semi-nonparametric maximum likelihood estimation. Econometrica: Journal of the Econometric Society **55**, 363–390 (1987)
[10] Hannan, E.J.: Rational transfer function approximation. Statistical Science **2**, 135–151 (1987)
[11] He, X.: Quantile curves without crossing. The American Statistician **51**, 186–192 (1997)
[12] Jin, Z., Lin, D.Y., Wei, L.J., Ying, Z.: Rank-based inference for accelerated failure time model. Biometrika **90**, 341–353 (2003)
[13] Jin, Z., Ying, Z., Wei, L.J.: A simple resampling method by perturbing the minimand. Biometrika **88**, 381–390 (2001)
[14] Koenker, R.: Quantile Regression. First Edition, Cambridge University Press (2005)
[15] Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**, 33–50 (1978)

[16] The Jackson Laboratory: Multi-system analysis of physiology on 7 inbred strains of mice. MPD:Jaxwest1. Mouse Phenome Database web site, The Jackson Laboratory. Bar Harbor, Maine USA, http://phenome.jax.org. Accessed 23 October 2012
[17] Li, T., Vuong, Q.: Nonparametric estimation of the measurement error model using multiple indicators. Journal of Multivariate Analysis **65**, 139–165 (1998)
[18] Matine, P., van Nes, S., Vanhoutte, E., Bakkers, M., van Doorn, P., Merkies, I., Faber, C.: Revised normative values fro grip strength with the Jamar dynamometer. Journal of Peripheral Nervous System **16**, 47–50 (2011)
[19] Molenaar, H.M.T., Selles, R.W., Zuidam, J.M., Willemsen, S.P., Stam, H.J., Hovius, S.E.R.: Growth diagrams for grip strength in children. Clin. Orthop. **468**, 217–223 (2010)
[20] Parzen, M.I., Wei, L.J., Ying, Z.: A resampling method basd on pivotal estimating functions. Biometrika **81**, 341–350 (1994)
[21] Reich,B.J., Bondell,H.D., Wang,H.J.: Flexible Bayesian quantile regression for independent and clustered data. Biostatistics **11**, 337–352 (2010)
[22] Schechtman, E., Spiegelman, C.: Mitigating the effect of measurement errors in quantile estimation. Statistics & Probability Letters **77**, 514–524 (2007)
[23] Staudenmayer, J., Ruppert, D., Buonaccorsi, J.P.: Density estimation in the presence of heteroscedastic measurement error. Journal of the American Statistical Association **103**, 726–736 (2008)
[24] Takada,T.: Asymptotic and qualitative performance of non-parametric density estimators: a comparative study. Econometrics Journal **11**, 573–592 (2008)
[25] Tooze, J.A., Kipnis, V., Buckman, D.W., Carroll, R.J., Freedman, L.S., Guenther, P.M., Krebs-Smith, S.M., Subar, A.F., Dodd, K.W.: A mixed-effects model approach for estimating the distribution of usual intake of nutrients: The NCI method. Stat. Med. **29**, 2857–2868 (2010)
[26] Variyam, J.N.: Factors affecting the macronutrient intake of U.S. adults. Electronic Report from the Economic Research Service (2003)
[27] Wang, H.J., Zhou, X.: Estimation of the retransformed conditional mean in health care cost studies. Biometrika **97**, 147–158 (2010)
[28] Zhang, D., Davidian, M.: Linear mixed models with flexible distributions of random effects for longitudinal data. Biometrics **57**, 795–802 (2001)

Hierarchical Bayesian Analysis of Repeated Binary Data with Missing Covariates

Fang Yu, Ming-Hui Chen, Lan Huang, and Gregory J. Anderson

Abstract Missing covariates are a common problem in many biomedical and environmental studies. In this chapter, we develop a hierarchical Bayesian method for analyzing data with repeated binary responses over time and time-dependent missing covariates. The fitted model consists of two parts: a generalized linear mixed probit regression model for the repeated binary responses and a joint model to incorporate information from different sources for time-dependent missing covariates. A Gibbs sampling algorithm is developed for carrying out posterior computation. The importance of the covariates is assessed via the deviance information criterion. We revisit the real plant dataset considered by Huang et al. (2008) and use it to illustrate the proposed methodology. The results from the proposed methods are compared with those in Huang et al. (2008). Similar top models and estimates of model parameters are obtained by both methods.

F. Yu (✉)
Department of Biostatistics, University of Nebraska Medical Center, Omaha, NE 68198-4375, USA
e-mail: fangyu@unmc.edu

M.-H. Chen
Department of Statistics, University of Connecticut, 215 Glenbrook Road U-4120, Storr, CT 06269, USA
e-mail: ming-hui.chen@uconn.edu

L. Huang
Office of Statistics, CDER, FDA, Silver Spring MD 20993, USA
e-mail: lan.huang@fda.hhs.gov

G.J. Anderson
Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, U-3043, Storrs, CT 06269, USA
e-mail: gregory.anderson@uconn.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_25,

1 Introduction

Missing covariates may occur due to the unavailability of covariate measurements, survey non-response, or data loss. A simple approach is to analyze the data that are fully observed, known as the complete case (CC) approach. When the data contain a large proportion of missing values in covariates, the CC approach may lead to biased and inefficient estimates of the model parameters. In addition, the subjects with missing covariates may differ systematically from those with completely observed covariates in terms of their associations with the outcome variable. Thus, it is important to incorporate the missing covariates in data analysis.

There is a rich literature on the methodological development for regression analyses with missing covariate data. Little (1992) reviewed these methods for multivariate normal regression models, while Ibrahim et al. (2005) presented a comprehensive review of missing data methods for generalized linear models. One of the major challenges in analyzing data with missing covariates is to model the distributions of missing covariates. Lipsitz and Ibrahim (1996) proposed a sequence of one-dimensional conditional distributions to model missing covariates. Huang et al. (2008) developed a maximum likelihood method to incorporate information from other data resource to model the time-dependent missing covariates and showed that their approach improved model fit on the repeated binary outcome variable. In this chapter, we develop a full Bayesian method to incorporate the third data resource for modeling time-dependent missing covariates and fit a probit regression model for the repeated binary response variable.

To illustrate the method proposed in this chapter, we consider the data analyzed in Huang et al. (2008) with a focus on the disruption of alternative flowering pattern of American basswood trees (*Tilia americana*). The alternative flowering pattern disrupts when the change in the flowing index (i.e., decreasing) between two consequent years fails to follow a change in opposite direction (i.e., increasing) in the following year. The flowering intensity data were collected annually on 24 trees over a 29-year period from 1974 to 2002. The binary outcome variable was defined with value 1 when the tree deviated from our expectation of an increasing value of the flowering intensity and 0 otherwise. The covariates were from two datasets: (1) the yearly defoliation data recording the defoliation by gypsy moths (*Lymantria dispar*) on each tree from 1973 to 2002 and (2) the monthly weather conditions including temperature and precipitation recorded as the "Departure from Normal Monthly Temperature" (denoted as "Temp") and "Departure from Normal Monthly Precipitation" (denoted as "Ppt") from the weather station in Storrs, Connecticut, the general area where the studied trees are located. The defoliation data were directly observed, while more than half weather data (including temperature and precipitation) were missing. To estimate the missing weather values, the neighboring stations within 25 miles of the local weather stations were identified, and the weather data from these neighboring stations were incorporated in the study.

The remainder of this chapter is organized as follows. Section 2 presents the Bayesian probit regression model for repeated binary responses. Section 3 develops

the model for missing weather conditions. In Sect. 4, the deviance information criterion (DIC) is derived for identifying important defoliation and/or weather condition factors associated with the cyclical pattern of flowering. A detailed analysis of the flower intensity data is given in Sect. 5. We conclude the chapter with brief remarks in Sect. 6.

2 The Model for Outcome Variable

2.1 Bayesian Probit Regression

Let y_{it}, $t = 1, \cdots, T_i$ denote the binary response variables over time of plant i for $i = 1, \cdots, n$. The response variable may be associated with both environmental covariates (i.e., weather) and plant-specific covariates (i.e., disease, pest). The environmental covariates often are time-dependent and shared by all plants at the same location. Missing values may exist in the environmental covariates. The plant-specific covariates are measured when collecting the response variable and, hence, they are often completely observed. Let $\mathbf{x}_t = (x_{t1}, \cdots, x_{tp})'$ be a p-dimensional vector for the environmental covariates and $\mathbf{z}_{it} = (z_{it1}, \cdots, z_{itq})'$ be a q-dimensional vector of plant-specific covariates at time t. In our real data example, $y_{it} = 1$ when the flower intensity of tree i at time t shows deviation from the alternating pattern for an expectation of an increasing value of the index and $y_{it} = 0$ otherwise. The covariates $\mathbf{x}_t$ contain the temperature and precipitation of year t observed from the local weather station, and the $\mathbf{z}_{it}$ contain the information of defoliation by gypsy moths for tree i.

The Bayesian probit regression is used to connect the response variable and the covariates. First, the latent variable ϖ_{it} is introduced to represent the real observed difference in flower intensities between adjacent years $t-1$ and t on the tree i. When $\varpi_{it} <= 0$, the flower intensities increase, which does not show a deviation from the expectation of an increasing value, and consequently $y_{it} = 0$. When $\varpi_{it} > 0$, there is a deviation from the alternating pattern and $y_{it} = 1$. Then the outcome variables y_{it} are independently distributed and the likelihood function of y_{it} is given by $L(\varpi_{it}|y_{it}) = (1\{\varpi_{it} \le 0\})^{1-y_{it}}(1\{\varpi_{it} > 0\})^{y_{it}}$, where $1\{A\}$ is the indicator function such that $1\{A\} = 1$ if A is true and 0 otherwise. We model ϖ_{it} using covariates $\mathbf{x}_t$ and $\mathbf{z}_{it}$ with

$$\varpi_{it} = \mu + b_i + \mathbf{x}_t'\beta + \mathbf{z}_{it}'\gamma + \varepsilon_{it}, \forall i,\, t, \tag{1}$$

where μ is the overall mean, b_i is the random plant effect, $\beta' = (\beta_1, \cdots, \beta_p)$ and $\gamma' = (\gamma_1, \cdots, \gamma_q)$ correspond to the effects of covariates $\mathbf{x}_t$ and $\mathbf{z}_{it}$ on ϖ_{it}, respectively, ε_{it} represents the remaining part of ϖ_{it} unexplained by these covariates. We assume $\varepsilon_{it} \sim N(0, \sigma^2)$ independently with $\sigma^2 = 1$ to ensure model identifiability.

Note that y_{it} may be missing at a given time t. Assume that y_{it} is missing at random (MAR). Then, we can marginalize the missing data of y_{it} and apply the

model in (1) only for the observed data of y_{it} using the covaraites X_t and Z_{it} from the same plant i of the same year t. Let δ_{it} be the indicator such that $\delta_{it} = 0$ if y_{it} is missing and 1 otherwise. Letting $\Theta_y = \{\mu, b_1, \cdots, b_n, \beta, \gamma\}$, the likelihood of the outcome variable y_{it} can be written as

$$\begin{aligned} & L(\Theta_y | y_{it}, \mathrm{Z}_{it}, \mathrm{X}_t, \delta_{it}) \\ & = \{1 - \Phi(\mu + b_i + \mathrm{X}_t'\beta + \mathrm{Z}_{it}'\gamma)\}^{(1-y_{it})} \{\Phi(\mu + b_i + \mathrm{X}_t'\beta + \mathrm{Z}_{it}'\gamma)\}^{y_{it}}, \end{aligned} \tag{2}$$

when $\delta_{it} = 1$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

In matrix form, (1) can be rewritten as $\varpi_i = \Delta_i \times \Theta + \varepsilon_i$ for $i = 1, \cdots, n$, where $\varpi_i = [\varpi_{i1}, \cdots, \varpi_{iT_i}]'$, $\Delta_i = [j_{T_i}, B_i, \mathrm{X}_i, \mathrm{Z}_i]$, j_{T_i} is the column vector containing T_i ones, B_i is a $T_i \times n$ matrix with all columns set to zero values except that the i^{th} column equals j_{T_i}, X_i is a $T_i \times p$ matrix with each row containing the corresponding X_t, and Z_i a $T_i \times q$ matrix with its t^{th} row equal to Z_{it}. Note that X_i is defined with index i to distinguish the plants as the outcome variable may not be observed in the same years for different plants. We further write $\Theta = [\mu, b_1, \cdots, b_n, \beta, \gamma]'$ and $\varepsilon_i = [\varepsilon_{i1}, \cdots, \varepsilon_{iT_i}]'$.

2.2 Prior Specification

Conjugate priors are specified for each parameter defined in (1). Specifically, we assume $\mu \sim N(0, \sigma_0^2)$, $b_i \sim N(0, \sigma_b^2)$, $\beta \sim N(0, c_0\sigma_1^2(\mathrm{X}'\mathrm{X})^{-1})$, and $\gamma \sim N(0, c_0\sigma_1^2(\mathrm{Z}'\mathrm{Z})^{-1})$, where X of dimension $N \times p$ and Z of dimension $N \times q$ consist of row sub-matrices X_i and Z_i, $i = 1, \cdots, n$, respectively, N equal the total number of outcome variables with $N = \sum_i T_i$, and $c_0 > 0$ is prespecified. We take $c_0 = 10$ in this chapter. The inverse gamma priors are assumed for the variance parameters with $\sigma_0^2 \sim IG(\gamma_0, \delta_0)$, $\sigma_b^2 \sim IG(\gamma_b, \delta_b)$, and $\sigma_1^2 \sim IG(\gamma_1, \delta_1)$, where the $(\gamma_0, \gamma_b, \gamma_1)$ are shape parameters and $(\delta_0, \delta_b, \delta_1)$ are scale parameters.

2.3 Posterior Distribution and Computation

Let $\varpi = [\varpi_1, \ldots, \varpi_n]$ be a collection of all latent variables, and $\Theta_p = \{\gamma_0, \delta_0, \gamma_b, \delta_b, \gamma_1, \delta_1\}$. To develop an efficient Gibbs sampling algorithm, we integrate out all random plant effects b_i in (2) and sample the remaining coefficient parameters and variance parameters using the modified collapsed Gibbs sampling method (Chen et al. 2000). Define a new covariates matrix $\Delta_i^* = [j_{T_i}, \mathrm{X}_i, \mathrm{Z}_i]$ for all i, and a new coefficient vector $\Theta^* = [\mu, \beta, \gamma]'$. Given $(\varpi, \mathrm{X}, \mathrm{Z}, \sigma_0^2, \sigma_b^2, \sigma_1^2)$, $\Theta^* \sim N_{1+p+q}(\mu_\Theta, \Sigma_\Theta)$, where $\Sigma_\Theta^{-1} = \sum_{i=1}^n (\Delta_i^{*'} U_i \Delta_i^*) + (\Sigma_\Theta^0)^{-1}$, $U_i = (I_{T_i} + J_{T_i}\sigma_b^2)^{-1} =$

$I_{T_i} - \frac{\sigma_b^2}{T_i\sigma_b^2+1}J_{T_i}$, I_{T_i} is the $T_i \times T_i$ identity matrix, and J_{T_i} is a square matrix of dimension $T_i \times T_i$ containing all ones. The matrix Σ_Θ^0 is the variance matrix defined in the prior distribution for Θ^*, and $(\Sigma_\Theta^0)^{-1} = \text{diag}\{\sigma_0^{-2}, c_0^{-1}\sigma_1^{-2}\text{X}'\text{X}, c_0^{-1}\sigma_1^{-2}\text{Z}'\text{Z}\}$. The mean vector $\mu_\Theta = \Sigma_\Theta^{-1} \times \sum_{i=1}^n (\Delta_i^{*'}\varpi_i)$. The posterior distributions of σ_0^2 and σ_1^2 in (1) are: $\sigma_0^2|\mu,\gamma_0,\delta_0 \sim IG(\gamma_0 + 1/2, \delta_0 + \mu^2/2)$ and $\sigma_1^2|\text{X},\text{Z},\beta,\gamma,\gamma_1,\delta_1 \sim IG(\gamma_1 + (p+q)/2, \delta_1 + (\beta'\text{X}'\text{X}\beta + \gamma'\text{Z}'\text{Z}\gamma)/(2c_0))$. Thus, sampling Θ^*, σ_0^2 and σ_1^2 is straightforward. The posterior distribution of σ_b^2 is given by

$$f(\sigma_b^2|\varpi,\text{X},\text{Z},\mu,\beta,\gamma,\delta_b,\gamma_b) \propto (\sigma_b^2)^{-\gamma_b-1}\exp(-\delta_b/\sigma_b^2)$$
$$\times (1+T_i\sigma_b^2)^{-\frac{1}{2}}\exp\left\{\frac{1}{2}\sum_{i=1}^n \frac{\sigma_b^2}{T_i\sigma_b^2+1}[\sum_{t=1}^{T_i}(\varpi_{it}-\mu-\text{X}_t'\beta-\text{Z}_{it}'\gamma)]^2\right\}. \quad (3)$$

The localized Metropolis algorithm (Chen et al. 2000) is used to sample σ_b^2 from (3). The latent vector ϖ_i follows a truncated multivariate normal distribution with $\pi(\varpi_i|\mu,\beta,\gamma,\text{X}_i,\text{Z}_i,\sigma_b^2) \propto f_{MN}(\varpi_i|\Delta_i^*\Theta, U_i^{-1})\prod_{t=1}^{T_i}[(1\{\varpi_{it} \le 0\})^{1-y_{it}}(1 - 1\{\varpi_{it} \le 0\})^{y_{it}}]$, in which $f_{MN}(\cdot|C,D)$ is the multivariate normal probability density function of mean vector C and variance matrix D. We sample ϖ_{it}, $t = 1,\cdots,T_i$ iteratively from a truncated normal distribution of $f_N((\varpi_{it}-\mu_{\varpi_{it}})/(\sigma_{\varpi_{it}})1\{w_{it} \le -\mu_{\varpi_{it}}/\sigma_{\varpi_{it}}\}$ if $y_{it} = 0$ and $f_N((\varpi_{it}-\mu_{\varpi_{it}})/\sigma_{\varpi_{it}})1\{\varpi_{it} > -\mu_{\varpi_{it}}/\sigma_{\varpi_{it}}\}$ if $y_{it} = 1$, where $\mu_{\varpi_{it}} = \frac{\sigma_b^2}{(T_i-1)\sigma_b^2+1}\sum_{t_1=1,t_1\neq t}^{T_i}(\varpi_{it_1} - \mu - \text{X}_{t_1}'\beta - \text{Z}_{it_1}'\gamma) + \mu + \text{X}_t'\beta + \text{Z}_{it}'\gamma$ and $\sigma_{\varpi_{it}}^2 = \frac{\sigma_b^2}{(T_i-1)\sigma_b^2+1} + 1$. The algorithm of Geweke (1991) is used to sample ϖ_{it}.

3 The Model for Missing Covariates

We only need to model the weather related time-dependent covariates x_t's since Z_{it}'s are completely observed. We define the observed covariates at time t as $\text{X}_{t,obs}$, the missing covariates as $\text{X}_{t,mis}$, and the indicator for missing with $r_{tj} = 1$ when x_{tj} is available and 0 otherwise. Assuming that the missing covariates x_{tj}'s are MAR, we only need to model X_t but not the missing data indicators (Ibrahim et al. 2005)). We modify the approach proposed by Huang et al. (2008) in the Bayesian framework to model the missing covariates.

Note that the information on covariates was recorded in neighboring stations, and the weather covariates among neighboring stations are expected to share similar patterns. We incorporate three sources of information when modeling the missing weather conditions (1) the weather information from neighboring stations; (2) the weather information from the previous months; and (3) the weather information from the previous year. Let w_{tjk}, $t = 1,\cdots,T$, $j = 1,\cdots,p$, $k = 1,\cdots,K$ be the observation from the k^{th} neighboring stations for the j^{th} weather variable of time t.

The $x_{tj} - w_{tjk}$ measures the deviance between x_{tj} from the local station and the corresponding variable w_{tjk} for $k = 1, \ldots, K$. We assume

$$x_{tj} - w_{tjk} = \mu_{tj} + \phi_{jk} + \varepsilon_{tjk}, \tag{4}$$

where μ_{tj} measures the effect of time and other covariates on the j^{th} covariate, ϕ_{jk} measures the effect of the k^{th} neighboring stations on the j^{th} weather variable, ε_{tjk} is a random error. We further assume $\varepsilon_{tj} = (\varepsilon_{tj1}, \varepsilon_{tj2}, \ldots, \varepsilon_{tjK})' \sim N(0, \Sigma_j)$, where Σ_j is a $K \times K$ covariance matrix, and ε_{tj}'s are independent for $t = 1, \ldots, T$ and $j = 1, \ldots, p$. In (4), μ_{tj} can be written into a linear function of the same covariate from the previous year denoted as $x_{t,j-1}$ and the values of other covariates at the same year including $x_{t,j-1}, \cdots, x_{t1}$ as follows:

$$\mu_{tj} = \xi_{j0} x_{t-1,j} + \xi_{j1} x_{t,j-1} + \cdots + \xi_{j,j-1} x_{t1}, j = 1, \cdots, p. \tag{5}$$

Depending on the characteristics of the covariates, we may just use a subset of the weather information from the previous year or months for μ_{tj}. Since a weak correlation was observed between the values of $x_{tj} - w_{tjk}$'s for different neighboring stations, we simply assume zero correlations between $(x_{tj} - w_{tjk})$'s for $k = 1, \cdots, K$. Accordingly, the variance covariance matrix Σ_j for ε_{tj} in (4) is written as $\Sigma_j = \text{diag}(\sigma^2_{j1}, \sigma^2_{j2}, \ldots, \sigma^2_{jK})$, where the σ^2_{jk} measures the conditional variance of $x_{tj} - w_{tjk}$ given μ_{tj}.

Since w_{tjk}'s may have a large percentage of missing values, for simplicity, we only include those observed w_{tjk}'s in (4) for modeling X_{tj}'s. Define $\zeta_{tjk} = 1$ to indicate that w_{tjk} is observed and 0 otherwise. Also let $D = \{y_{it}, \delta_{it}, \mathrm{Z}_{it}, \mathrm{X}_{t,mis}, \mathrm{X}_{t,obs}, \ \zeta_{tjk} w_{tjk}, b_i, 1 \le i \le n, 1 \le t \le T, 1 \le k \le K\}$ denote the complete data. Then, the complete-data likelihood function is given by $L(\Theta_y, \sigma_b, \xi, \phi, \Sigma | D) \propto \prod_{i=1}^{n} \{ \big[\prod_{t=1}^{T} \big(1 - \Phi(\mu + b_i + \mathrm{X}_t'\beta + \mathrm{Z}_{it}'\gamma)\big)^{(1-y_{it})\delta_{it}} \big(\Phi(\mu + b_i + \mathrm{X}_t'\beta + \mathrm{Z}_{it}'\gamma)\big)^{y_{it}\delta_{it}} \big] (\sigma_b^2)^{-1/2} \ \exp(-b_i^2/\ (2\sigma_b^2)) \} \times \big[\prod_{t=1}^{T} \prod_{k=1}^{K} \prod_{j=1}^{p} \sigma_{jk}^{-\zeta_{jtk}} \exp\{-\zeta_{jtk}(x_{tj} - w_{tjk} - \mu_{tj} - \phi_{jk})^2/(2\sigma^2_{jk})\} \big]$, where μ_{tj} is defined in (5), $\xi = (\xi_{j0}, \xi_{j1}, \xi_{j2} \ldots \xi_{j,j-1}, 1 \le j \le p)'$, $\phi = (\phi_{jk}, 1 \le j \le p, 1 \le k \le K)'$, and $\Sigma = (\sigma^2_{jk}, 1 \le k \le K, 1 \le j \le p)'$.

3.1 Prior Specification

We assume a conjugate prior for each parameter in Θ_x. Specifically, $\phi_{jk} \sim N(0, \sigma^2_{\phi_j})$, $\xi_{jl} \sim N(0, \sigma^2_{\xi_j})$ for $l = 1, \cdots, L_j$, where L_j equals the number of covaraites in defining μ_{tj} in (5). We further assume $\sigma^2_{\phi_j} \sim IG(\nu_\phi, \rho_\phi)$ and $\sigma^2_{\xi_j} \sim IG(\nu_\xi, \rho_\xi)$. For σ^2_{jk}, we assume that $\sigma^2_{jk} \sim IG(\nu_{s_j}, \rho_{s_j})$ and their parameters ρ_{s_j} independently follow gamma distributions with $\rho_{s_j} \sim G(\nu_{s_0}, \rho_{s_0})$.

3.2 Posterior Distribution and Computation

The posterior distribution of x_{tj} is $N(\mu^*_{x_{tj}}, \sigma^{*2}_{x_{tj}})$, where $\mu^*_{x_{tj}} = \sigma^{*2}_{x_{tj}}(\frac{\mu_{x_{tj}}}{\sigma^2_{x_{tj}}} + \sum_{i=1}^{n} \delta_{it}\beta_j(w_{it} - \mu - b_i - Z_{it}\gamma - \sum_{j1\neq j} x_{tj_1}\beta_{j_1}))$, $\sigma^{*2}_{x_{tj}} = (\sigma^{-2}_{x_{tj}} + \sum_{i=1}^{n} \delta_{it}\beta_j^2)^{-1}$, and $\sigma^{-2}_{x_{tj}} = \sum_{k=1}^{K} \zeta_{tjk}/\sigma^2_{jk}$. We use the Gibbs sampling algorithm to sample the missing covariates x_{tj}. Similarly, the posterior distribution of ϕ_{jk} given X_t, μ_{tj}'s, w_{tjk}'s, σ^2_{jk}'s, $\sigma^2_{\phi_j}$'s and the posterior distribution of ξ_{jl}'s given X_t, ϕ_{jk}'s, w_{tjk}'s, σ^2_{jk}'s, $\sigma^2_{\xi_j}$'s follow normal distributions. Since multiple parameters ϕ_{jk}'s and ξ_{jl}'s are involved in the model for missing covariates, it is computationally extensive to sample each parameter iteratively. Instead we sample ϕ_{jk}'s and ξ_{jl}'s jointly from their multivariate normal posterior distribution given X_t, w_{tjk}'s, σ^2_{jk}'s, $\sigma^2_{\phi_j}$'s and $\sigma^2_{\xi_j}$'s. The posterior distributions of $\sigma^2_{\phi_j}$ and $\sigma^2_{\xi_j}$ are given by $\sigma^2_{\phi_j} \sim IG(\nu_\phi + \frac{K}{2}, \rho_\phi + \frac{1}{2}\sum_{k=1}^{K}(\phi_{jk})^2)$ and $\sigma^2_{\xi_j} \sim IG(\nu_\xi + \frac{L_j}{2}, \rho_\xi + \frac{1}{2}\sum_{l=1}^{L_j}(\xi_{jl})^2)$. The variance parameters σ^2_{jk} have inverse gamma posterior distributions of $\sigma^2_{jk} \sim IG(\nu_{s_j} + \sum_{t=1}^{T}\frac{\zeta_{tjk}}{2}, \rho_{s_j} + \sum_{t=1}^{T}\frac{\zeta_{tjk}}{2}(x_{tj} - \mu_{tj} - w_{tjk} - \phi_{jk})^2)$, and their parameters ρ_{s_j} independently follow gamma posterior distributions of $\rho_j \sim G(\nu_{s_0} + K\nu_{s_j}, 1/(\rho_{s_0}^{-1} + \sum_{k=1}^{K}\sigma^{-2}_{jk}))$. Thus, sampling these parameters is straightforward.

4 DIC for Model Comparison

The Deviance Information Criterion (DIC) of Spiegelhalter et al. (2002) is used for identifying important subset of covariates. The models for outcome variables and covaraites involve parameters Θ_y and Θ_x. Accordingly, DIC can be calculated by $DIC = D(\overline{\Theta}_y, \overline{\Theta}_x) + 2p_D$, where the penalty term $p_D = \overline{D(\Theta_y, \Theta_x)} - D(\overline{\Theta}_y, \overline{\Theta}_x)$, and $\overline{\Theta}_y = E(\Theta_y | D_{obs})$. Since our primary interest is on selecting important subset of β and γ for fitting the outcome variables, we focus on the Bayesian probit regression model for the outcome variables. Then we consider X as a parameter, and have $\Theta^* = \{\Theta_y, X\}$, and DIC becomes $DIC = D(\overline{\Theta^*}) + 2p_D$, where $p_D = \overline{D(\Theta^*)} - D(\overline{\Theta^*})$. Based on (2) for the outcome variable, we define the deviance function $D(\Theta^*)$ as $-2\log L(\Theta_y, X_t | y_{it}, Z_{it}, \delta_{it})$.

Huang et al. (2005) developed a version of DIC for the generalized linear model with missing covariates. Note that the covaraites X's were involved in the outcome variable through term $X_t'\beta$. Instead of estimating X_t and β separately for calculating the DIC value, following the suggestion by Huang et al. (2005), we redefine our parameter set as $\Theta^\sharp = \{\mu, X'\beta, \gamma\}$ and calculate the DIC value. When the covariate X_t is observed then $\overline{X_t'\beta} = X_t' E(\beta | D)$, otherwise $\overline{X_t'\beta} = E(X_t'\beta | D)$. In addition, we can integrate out b_i from the likelihood function for the outcome variable in (2). As a result, we need to calculate product of cumulative distribution function of multivariate normal distributions via the algorithm proposed by Genz (1992).

5 Results

The proposed method is applied to analyze the real flower intensity data described in Sect. 1. Our primary interest is to evaluate the effects of important variables (including both weather variables and defoliation variables) on the pattern of flowering intensities. We consider the same 19 models discussed in Huang et al. (2008). We set $c_0 = 10$ and 0.01 for all shape and scale hyperparameters $\{\gamma_0, \delta_0, \gamma_b, \delta_b, \gamma_1, \delta_1\}$ in the model for the outcome variable and $\{\nu_\xi, \rho_\xi, \nu_\phi, \rho_\phi, \nu_{s_0}, \rho_{s_0}\}$ in the model for missing covariates. The DIC criterion measures are used to compare all 19 models. The models with smaller DIC values fit the data better. In the first 7 models listed in Table 1, the effects of temperature conditions, precipitation conditions, defoliation status, or their combinations are included. Note that the weather conditions from different seasons may affect the flowering intensities differently by influencing the initiation or development of the flowers. The monthly weather information is split into seasons. We focus on the 3 months before July or after July, when the flowering reaches its peak. We define the months of April to June of current year as "spring season," and the months of August to October of previous year as "fall season." In the remaining 12 models listed in Table 1, the weather information from particular seasons are used to identify the seasonal effects of weather conditions on flowering intensities. Table 1 shows the values of DIC and p_D for all 19 models. Five of these models have DIC value less than 160. Out of these 5 models, model 5 (containing

Table 1 The values of DIC for 19 models

			Bayesian	
Model	Definition	ML[a]	DIC	P_d
1	Defoliation (Defol-1, Defol)	10.85	166.52	3.20
2	Ppt (PptJul-1, …, PptJul)	15.62	166.91	10.21
3	Temp (TempJul-1, …, TempJul)	8.67	161.92	11.15
4	Defoliation & Ppt (Model 1 + Model 2)	10.14	162.77	11.06
5	Defoliation & Temp (Model 1 + Model 3)	4.25	157.29	11.99
6	Ppt & Temp (Model 2 + Model 3)	21.93	165.62	17.26
7	Defoliation, Ppt & Temp (Models 1 + 2 + 3)	22.51	162.01	17.40
8	Fall Ppt (PptAug-1, PptSept-1, PptOct-1)	30.96	183.21	2.01
9	Spring Ppt (PptApr, PptMay, PptJun)	12.64	168.27	4.02
10	Fall Temp (TempAug-1, TempSept-1, TempOct-1)	21.12	177.41	3.59
11	Spring Temp (TempApr, TempMay, TempJun)	22.84	177.95	3.73
12	Spring & Fall Ppt (Model 8 + Model 9)	10.03	165.28	6.24
13	Spring & Fall Temp (Model 10 + Model 11)	1.17	158.20	6.64
14	Fall Ppt & Temp (Model 8 + Model 10)	25.42	179.74	4.58
15	Spring Ppt & Temp (Model 9 + Model 11)	10.28	165.80	6.59
16	Spring & Fall Ppt & Defol-1 (Model 12 + Defol-1)	0	159.33	6.96
17	Spring & Fall Temp & Defol-1 (Model 13 + Defol-1)	1.05	155.55	7.10
18	Fall Ppt & Temp & Defol-1 (Model 14 + Defol-1)	7.27	171.34	6.31
19	Spring Ppt & Temp & Defol-1 (Model 15 + Defol-1)	1.51	159.19	7.15

[a]: Maximum Likelihood approach, $\mathrm{AIC}_Q - 1298.36$ were listed.

temperature and defoliation) contains more than twice number of covariates in other four models. Specifically, model 13 and model 17 include Spring and Fall temperature (versus the 13 monthly values for the full flowering cycle, as model 5), model 16 includes Spring and Fall precipitation, and model 19 includes Spring precipitation and temperature. Defoliation from the earlier year is also included in models 16, 17, and 19. The DIC values from all these 5 models are comparable. The big model has the small DIC value since the p_D value is relatively small compared to the complexity of the model. Though the DIC values may be in favor of a large model due to underestimation in the dimensional penalty, the DIC values do show that the smaller penalty is associated with a smaller model. Therefore, we select the four models containing small subsets of covariates with comparable DIC values. We note that these four models identified based on DIC also are selected by the AIC_Q in Huang et al. (2008).

Since model 13 is nested in model 17, we focus on the other three models of the four top models and assess the effects of the selected covariates on the flowering intensities. The posterior means, the posterior standard deviations (SDs), and the 95% highest posterior density (HPD) intervals were given in Table 2. In all the three models, the defoliation from previous year (Defol-1) is included in the model and its 95 % HPD interval is above zero in models 16 and 19, implying a positive effect, and high levels of defoliation are likely to cause a deviation in the alternating pattern when an increase in flowering intensity is expected. The precipitation from June (PptJun) is included in both models 16 and 19, showing significant negative effects in both models. Hence, the higher the precipitation in June of the current year, the less likely to observe a deviation in the increasing pattern of flowering intensity. The temperature also shows an important influence in the flowering intensity. In model 17, the temperature of August (TempAug-1) from the previous year and May (TempMay) of the current year has negative effects.

The estimated effects shown in Table 2 have biological meaning and are expected. Using defoliation as an example, the defoliation by gypsy moths in the previous year apparently reduces the ability of trees to reserve energy and prepare for the flowering in the current year. June is a critical month for flowering intensities. The precipitation in June of the current year will help the tree to develop leaves or flowers; hence, a high precipitation in these two months will promote flowering intensities. Similarly, it is expected that the higher temperature in August of the previous year and May of the current year, the more energy the tree can reserve for developing new leaves, and heavier flowering intensities will be observed.

Table 2 also presents the estimates of Huang et al. (2008) using the maximum likelihood (ML) method for models 16, 17, and 19. As the logistic regression model was used for outcome variable in Huang et al. (2008), the estimates of the regression coefficients are expected to share similar directions but have different magnitudes. From Table 2, we see some slight differences between these two methods in model 17 but not in models 16 and 19. In model 17, temperature in September of previous year shows a significant negative effect and defoliation of previous year shows a

Table 2 Posterior estimates of the parameters under three top models

		Maximum Likelihood			Bayesian		
Model	Variable	Estimate	SE	P-Value	Mean	SD	95% HPD
16	Intercept	-1.517	0.239	<0.001	-0.794	0.127	(-1.034, -0.543)
	PptAug-1	0.189	0.302	0.532	0.056	0.129	(-0.204, 0.307)
	PptSep-1	0.032	0.238	0.894	0.085	0.156	(-0.216, 0.390)
	PptOct-1	-0.310	0.271	0.252	-0.219	0.118	(-0.434, 0.033)
	PptApr	-0.048	0.290	0.870	0.150	0.124	(-0.084, 0.413)
	PptMay	-0.498	0.383	0.193	-0.052	0.127	(-0.302, 0.199)
	PptJun	-1.198	0.244	<0.001	-0.577	0.190	(-0.940, -0.202)
	Defol-1	1.230	0.319	<0.001	0.501	0.233	(0.069, 0.982)
	σ^2	0.200	0.313	–	0.004	0.022	(0.000, 0.014)
17	Intercept	-1.545	0.260	<0.001	-0.822	0.131	(-1.089, -0.575)
	TempAug-1	-0.803	0.363	0.027	-0.402	0.148	(-0.696, -0.113)
	TempSep-1	-0.671	0.248	0.007	-0.230	0.142	(-0.509, 0.050)
	TempOct-1	0.351	0.272	0.197	0.121	0.119	(-0.112, 0.347)
	TempApr	0.605	0.340	0.075	0.267	0.159	(-0.018, 0.601)
	TempMay	-1.026	0.320	0.001	-0.476	0.171	(-0.799, -0.124)
	TempJun	-0.144	0.315	0.649	0.038	0.148	(-0.238, 0.343)
	Defol-1	0.741	0.317	0.019	0.366	0.205	(-0.001, 0.806)
	σ^2	0.078	0.457	–	0.003	0.016	(0.000, 0.012)
19	Intercept	-1.546	0.243	<0.001	-0.792	0.128	(-1.053, -0.555)
	PptApr	-0.101	0.255	0.692	0.080	0.113	(-0.140, 0.293)
	PptMay	-0.627	0.423	0.138	-0.141	0.123	(-0.392, 0.082)
	PptJun	-1.154	0.235	<0.001	-0.443	0.191	(-0.837, -0.099)
	TempApr	0.308	0.267	0.249	0.107	0.138	(-0.162, 0.377)
	TempMay	-0.450	0.322	0.162	-0.214	0.129	(-0.469, 0.032)
	TempJun	-0.157	0.247	0.525	0.021	0.147	(-0.247, 0.328)
	Defol-1	1.217	0.313	<0.001	0.501	0.226	(0.085, 0.936)
	σ^2	0.021	0.446	–	0.003	0.019	(0.000, 0.015)

SE, Standard Error; HPD, highest posterior density interval

significant positive effect only in the ML method. However, the Bayesian method shows that both covariates have marginal effects.

Following Chen et al. (2000), we define the Bayesian latent residuals as $\varpi_{it}^* = (\varpi_{it} - \mu_{\varpi_{it}})/\sigma_{\varpi_{it}}$, where $\mu_{\varpi_{it}} = E(\varpi_{it}|D)$ and $\sigma_{\varpi_{it}}^2 = Var(\varpi_{it}|D)$ are the posterior mean and variance of ϖ_{it}. Accordingly we calculate the posterior probabilities $P(|\varpi_{it}^*| \geq 2|D)$ using the MCMC samples. Similarly, we estimate $E(y_{it,new}|D) = E[\Phi(\mu + b_i + X_{it}'\beta + Z_{it}'\gamma)|D]$ using the MCMC samples. Figure 1 shows the scatter plots of the posterior probabilities $P(|\varpi_{it}^*| \geq 2|D)$ versus $E(y_{it,new}|D)$ for the three top models to assess the model adequacy. The estimated probabilities $P(|\varpi_{it}^*| \geq 2|D)$ for all ϖ_{it}^* in Fig. 1 are less than 0.06, indicating that the three top models all fit the flower intensity data fairly well.

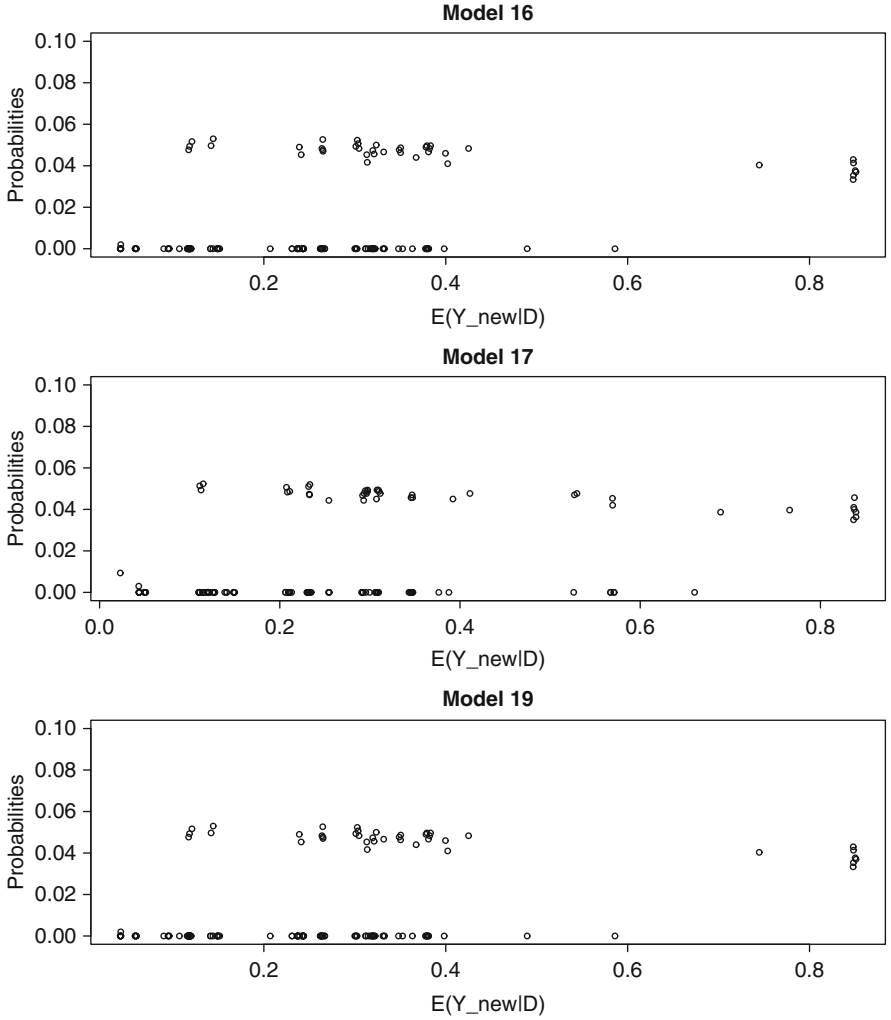


Fig. 1 Plots of posterior probabilities $P(|\varpi_{it}^* \geq 2|D)$ versus $E(y_{it,new}|D)$ for three top models.

6 Discussion

We have developed a useful Bayesian approach for modeling missing covariates in addition to the probit regression model for repeated binary responses. The DIC measure is derived to identify important covariates when data contain missing covariates. The results from the real data analysis on flowering intensity pattern of *Tilia americana* demonstrate that (1) the incorporation of additional information (i.e., weather information from neighboring weather stations) improves the model fit and (2) both DIC and AIC_Q select similar top models. We note that the Bayesian

and ML methods yield slightly different estimates of the model parameters. The difference between the Bayesian method and the ML method may be partially due to the relatively small sample size. As it is well known, the Bayesian estimates including the posterior standard deviations and 95 % HPD intervals are calculated without resorting to asymptotics while the ML method does require the asymptotic normality.

Acknowledgments Thanks to Dr. Paul Neal for analysis of some of the field data. We would also like to thank the editors for helpful comments and suggestions, which have led to an improved version of this chapter.

References

Chen, M.-H., Shao, Q.-M., Ibrahim, J.G.: Monte Carlo Methods in Bayesian Computation. Springer-Verlag, New York (2000)

Genz, A.: Numerical computation of multivariate normal probabilities. J. Comput. Graph. Statist. **1**, 141–149 (1992)

Geweke, J.: Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In: Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface. Fairfax Station, VA: Interface Foundation of North America Inc., pp. 571–578. (1991)

Huang, L., Chen, M.-H., Ibrahim, J.G.: Bayesian analysis for generalized linear models with nonignorably missing covaraites. Biometrics **61**, 767–780 (2005)

Huang, L., Chen, M.-H., Yu, F., Neal, P.R., Anderson, G.J.: On modeling repeated binary response and time-dependent missing covariates. J. Agric. Biol. Environ. Stat. **13**, 270–293 (2008)

Ibrahim, J.G., Chen, M.-H., Lipsitz, S.R., Herring, A.H.: Missing data methods in regression models. J. Amer. Statist. Assoc. **100**, 332–346 (2005)

Lipsitz, S.R, Ibrahim, J.G.: A conditional model for incomplete covariates in parametric regression models. Biometrika **83**, 916–922 (1996)

Little, R.J.A.: Regression with missing X's: A review. J. Amer. Statist. Assoc. **87**, 1227–1237 (1992)

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.: Bayesian measures of model complexity and fit (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. **64**, 583–639 (2002)

Part VIII
Multi–Regional Clinical Trials

Use of Random Effect Models in the Design and Analysis of Multi-regional Clinical Trials

Yuh-Jenn Wu, Te-Sheng Tan, Shein-Chung Chow, and Chin-Fu Hsiao

Abstract In recent years, global collaboration has become a commonly used strategy for new drug development. To accelerate the development process and shorten the approval time, the design of multi-regional clinical trials (MRCTs) incorporates subjects from many countries around the world under the same protocol. After showing the overall efficacy of a drug in all global regions, one can also simultaneously evaluate the possibility of applying the overall trial results to all regions and subsequently support drug registration in each of them. Several statistical methods have been proposed for the design and evaluation of MRCTs. Most of these approaches, however, assume a common variability of the primary endpoint across regions. In practice, this assumption may not be true due to differences across regions. In this paper, we use a random effect model for modeling heterogeneous variability across regions for the design and evaluation of MRCTs.

Y.-J. Wu
Department of Applied Mathematics, Chung Yuan Christian University, Chung Li, Taiwan

T.-S. Tan
Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, 35 Keyan Road, Zhunan Town, Miaoli County 350, Taiwan

S.-C. Chow
Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27705, USA

C.-F. Hsiao (✉)
Division of Clinical Trial Statistics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan Town, Miaoli County, Taiwan
e-mail: chinfu@nhri.org.tw

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_26,

1 Introduction

To speed up drug development and regulatory approval time, the design of multi-regional clinical trials (MRCTs) incorporates subjects from many countries around the world under the same protocol. After showing the overall efficacy of a drug in all global regions, one can also simultaneously evaluate the possibility of applying the overall trial results to all regions and subsequently support drug registration in each of them. Recent approaches for sample size determination in multi-regional trials developed by Kawai et al. (2008) and Ko et al. (2010) are all based on the assumption that the effect size is uniform across regions. In practice, it might be expected that there is a difference in treatment effect due to regional difference (e.g., ethnic difference). Hung et al. (2010) presented a number of useful statistical analysis tools for exploration of regional differences and a method that may be worth consideration in planning an MRCT. Chen et al. (2012) proposed a random effect model for heterogeneous treatment effect across regions for the design and evaluation of an MRCT. For those approaches, an equal variability of the primary endpoint across regions is usually assumed.

In this paper, we concern the case that the mean of response variable is fixed but the shift in the variance across regions is random. This assumption may not be trivial. In a population of patients with chronic pain, no group differences were evident for the STAI (State-Trait Anxiety Inventory) state anxiety scores, but the observed standard deviation from African-American population might be smaller than that from Caucasian populations (Edward et al. 2001). In subjects with diabetes mellitus, there was no difference in low density lipoprotein (LDL)/apolipoprotein (apo) B ratio between Asian Indians and Chinese, but there might exist differences in the observed standard deviation between two groups. In fact, the observed standard deviation from the Asian Indians was much larger than that observed from Chinese (Tan et al. 2001).

This paper is organized as follows. In Sect. 2, we evaluate the drug efficacy in overall region when unequal variability of the primary endpoint across regions is assumed in the MRCT. A method for sample size determination of the MRCT is also proposed in Sect. 3. Numerical examples are presented in Sect. 4 to illustrate the use of our method. Discussions are given in Sect. 5.

2 Statistical Hypothesis and Testing Statistic

For simplicity, we focus on the MRCT for comparing a test product and a placebo control based on a continuous efficacy endpoint. Suppose that there will be M regions participated in the MRCT. Let X_{ij} and Y_{ik} be efficacy responses for patients j and k receiving the test product and the placebo control, respectively, in the i^{th} region, $i=1,\ldots,M$, $j=1,\ldots, n_i^{\mathrm{T}}$, and $k=1,\ldots, n_i^{\mathrm{C}}$. In this paper, we assume a common treatment effect but an unequal variability of the efficacy response across regions. Subsequently, we can assume that

$$X_{ij} \sim N(\mu_T, \sigma_{Ti}^2), \text{ and } Y_{ik} \sim N(\mu_C, \sigma_{Ci}^2), \tag{1}$$

where $N(\mu, \sigma^2)$ represents the normal distribution with mean μ and σ^2. To address the issue of heterogeneous variability across regions, similar to the approach developed by Lu et al. (2010), we can let

$$\sigma_{Ti} = \frac{\sigma_T}{\tau_{Ti}}, \ \sigma_{Ci} = \frac{\sigma_C}{\tau_{Ci}},$$

and assume that

$$\tau_{Ti} \sim Gamma(\nu_T, \lambda_T \nu_T), \quad \text{and} \quad \tau_{Ci} \sim Gamma(\nu_C, \lambda_C \nu_C), \tag{2}$$

where $\lambda_T > 0$ and $\lambda_C > 0$. Here, τ_{Ti} and τ_{Ci} can be thought of as the shift in the variance across regions. Following this design, μ_T, μ_C, λ_T, λ_C, ν_T, ν_C, σ_T^2 and σ_C^2 are the unknown parameters.

For the treatment group, in addition to observe $X = \{x_{ij}, j = 1, \ldots, n_i^T; i = 1, \ldots, M\}$ from all regions, we can consider $\tau_T = \{\tau_{T1}, \ldots, \tau_{TM}\}$ as the latent data. Therefore, the log-likelihood function can be written as

$$\begin{aligned}
&L(u_T, \sigma_T^2, \nu_T, \lambda_T | X, \tau_T) \\
&\propto L(X | u_T, \sigma_T^2, \tau_T) + L(\tau_T | \nu_T, \lambda_T) \\
&= -n^T \log \sigma_T + \sum_{i=1}^{M} \tfrac{n_i^T}{2} \log \tau_{Ti} - \tfrac{1}{2\sigma_T^2} \sum_{i=1}^{M} \sum_{j=1}^{n_i^T} \tau_{Ti} (x_{ij} - u_T)^2 + \nu_T M \log(\lambda_T \nu_T) - \\
&(M+1) \log \Gamma(\nu_T) + (\nu_T - 1) \sum_{i=1}^{M} \log \tau_{Ti} - (\lambda_T \nu_T) \sum_{i=1}^{M} \tau_{Ti},
\end{aligned} \tag{3}$$

where $n^T = \sum_{i=1}^{M} n_i^T$. Since there doesn't exist closed-form solution to the likelihood equations, we will use the expected maximum (EM) algorithm to obtain the maximum likelihood estimates of u_T, σ_T^2, ν_T, and λ_T.

The EM algorithm is comprised of two steps, namely E-step and M-step. The E-step computes the conditional expectation of the complete-data sufficient statistics given the observed values and the current estimates $\Omega_T^{(t)} = \{X, u_T^{(t)}, \sigma_T^{(t)}, \nu_T^{(t)}, \lambda_T^{(t)}\}$ and the M-step maximizes the expected $L(X | u_T, \sigma_T^2, \tau_T)$ over (u_T, σ_T^2) and the expected $L(\tau_T | \nu_T, \lambda_T)$ over (ν_T, λ_T) to derive the local maximum values. The EM algorithm proceeds as follows.

E-Step: By conditioning on the current estimates $\Omega_T^{(t)} = \{X, u_T^{(t)}, \sigma_T^{(t)}, \nu_T^{(t)}, \lambda_T^{(t)}\}$, we have

$$\tau_{Ti} \sim \text{Gamma}(\nu_T^{(t)} + \frac{n_i^T}{2}, \lambda_T^{(t)} \nu_T^{(t)} + \sum_{j=1}^{n_i^T} \frac{(x_{ij} - u_T^{(t)})^2}{2(\sigma_T^{(t)})^2}).$$

The expectations of the relevant statistics are given by

$$E[\tau_{Ti}|\Omega_T^{(t)}] = \frac{2\nu_T^{(t)}(\sigma_T^{(t)})^2 + n_i^T(\sigma_T^{(t)})^2}{2\lambda_T^{(t)}\nu_T^{(t)}(\sigma_T^{(t)})^2 + \sum_{j=1}^{n_i^T}(x_{ij} - u_T^{(t)})^2},$$

and

$$E[\log\tau_{Ti}|\Omega_T^{(t)}] = \phi(\nu_T^{(t)} + \frac{n_i^T}{2}) - \log(\lambda_T^{(t)}\nu_T^{(t)} + \sum_{j=1}^{n_i^T}\frac{(x_{ij} - u_T^{(t)})^2}{2(\sigma_T^{(t)})^2}),$$

where $\phi(y) = \frac{\Gamma'(y)}{\Gamma(y)}$ represents the digamma function. Given the current estimates $\Omega_T^{(t)} = \{X, u_T^{(t)}, \sigma_T^{(t)}, \nu_T^{(t)}, \lambda_T^{(t)}\}$, let

$$V_{T1i}{}^{(t)} = E[\tau_{Ti}|\Omega_T^{(t)}] \text{ and } V_{T2i}{}^{(t)} = E[\log\tau_{Ti}|\Omega_T^{(t)}].$$

We can obtain the expectation of the log-likelihood function:

$$\begin{aligned}
&Q(u_T, \sigma_T^2, \nu_T, \lambda_T|\Omega_T^{(t)}) \\
&= -n^T\log\sigma_T + \sum_{i=1}^{M}\frac{n_i^T}{2}V_{T2i}{}^{(t)} - \frac{1}{2\sigma_T^2}\sum_{i=1}^{M}\sum_{j=1}^{n_i^T}V_{T1i}{}^{(t)}(x_{ij} - u_T)^2 + \nu_T M\log(\lambda_T\nu_T) \\
&-M\log\Gamma(\nu_T) + (\nu_T - 1)\sum_{i=1}^{M}V_{T2i}{}^{(t)} - (\lambda_T\nu_T)\sum_{i=1}^{M}V_{T1i}{}^{(t)}.
\end{aligned} \tag{4}$$

M-step: We can maximize the function (4) by

$$\begin{aligned}
\frac{\partial Q}{\partial u_T} &= \frac{1}{\sigma_T^2}\sum_{i=1}^{M}\sum_{j=1}^{n_i^T}V_{T1i}{}^{(t)}(x_{ij} - u_T) = 0, \\
\frac{\partial Q}{\partial \sigma_T^2} &= -\frac{n_T}{\sigma_T^2} + \frac{1}{\sigma_T^3}\sum_{i=1}^{M}\sum_{j=1}^{n_i^T}V_{T1i}{}^{(t)}(x_{ij} - u_T)^2 = 0, \\
\frac{\partial Q}{\partial \lambda_T} &= \frac{\nu_T M}{\lambda_T} - \nu_T\sum_{i=1}^{M}V_{T1i}{}^{(t)} = 0,
\end{aligned}$$

and

$$\frac{\partial Q}{\partial \nu_T} = M\log(\lambda_T\nu_T) + M - M\phi(\nu_T) + \sum_{i=1}^{M}V_{T2i}{}^{(t)} - \lambda_T\sum_{i=1}^{M}V_{T1i}{}^{(t)} = 0. \tag{5}$$

The iterations of the estimates of parameters $u_T, \sigma_T^2,$ and λ_T are given by

$$\hat{u}_T^{(t+1)} = \frac{\sum_{i=1}^{M} V_{T1i}{}^{(t)} (\sum_{j=1}^{n_i^T} x_{ij})}{\sum_{i=1}^{M} n_i^T V_{T1i}{}^{(t)}},$$
$$(\hat{\sigma}_T^{(t+1)})^2 = \frac{1}{n_T} \sum_{i=1}^{M} \sum_{j=1}^{n_i^T} V_{T1i}{}^{(t)} (x_{ij} - u_T^{(t+1)})^2, \quad (6)$$
$$\text{and}$$
$$\hat{\lambda}_T^{(t+1)} = \frac{M}{\sum_{i=1}^{M} V_{T1i}{}^{(t)}}.$$

In addition, $\nu_T^{(t+1)}$ can be derived by finding the solution to Eq. (5) by Newton method, i.e.,

$$\hat{\nu}_T^{(t+1)} = \nu_T^{(t)} - \frac{\log(\lambda_T^{(t+1)} \nu_T^{(t)}) - \phi(\nu_T^{(t)}) + \frac{1}{M} \sum_{i=1}^{M} V_{T2i}{}^{(t)} - \frac{\lambda_T^{(t+1)}}{M} \sum_{i=1}^{M} V_{T1i}{}^{(t)} + 1}{\frac{1}{\nu_T^{(t)}} - \frac{\phi'(\nu_T^{(t)})}{2}}. \quad (7)$$

Proceeding similarly, we can use the EM algorithm to derive the maximum likelihood estimates of u_C, σ_C^2, ν_C and λ_C.

Assume that the EM iteration converges after the $(t+1)^{th}$ loop. By Theorem 1 in Lu et al. (2010), we can approximately derive that

$$E(u_T^{(t+1)}) \approx u_T,$$
$$E(u_C^{(t+1)}) \approx u_C,$$
$$Var(u_T^{(t+1)}) \approx \frac{\lambda_T \sigma_T^2 \nu_T \sum_{i=1}^{M} n_i^T (V_{T1i}{}^{(t)})^2}{(\nu_T - 1)(\sum_{i=1}^{M} n_i^T V_{T1i}{}^{(t)})^2},$$

and

$$Var(u_C^{(t+1)}) \approx \frac{\lambda_C \sigma_C^2 \nu_C \sum_{i=1}^{M} n_i^C (V_{C1i}{}^{(t)})^2}{(\nu_C - 1)(\sum_{i=1}^{M} n_i^C V_{C1i}{}^{(t)})^2}.$$

Let $\tilde{\mu}_T$ and $\tilde{\mu}_C$ be the approximate MLEs. Subsequently, we have

$$\tilde{\mu}_T - \tilde{\mu}_C \sim N(\mu_T - \mu_C, \sigma_p^2),$$

where

$$\sigma_p^2 = \frac{\lambda_T \sigma_T^2 \nu_T \sum_{i=1}^{M} n_i^T \left(V_{T1i}^{(t)}\right)^2}{(\nu_T - 1)\left(\sum_{i=1}^{M} n_i^T V_{T1i}^{(t)}\right)^2} + \frac{\lambda_C \sigma_C^2 \nu_C \sum_{i=1}^{M} n_i^C \left(V_{C1i}^{(t)}\right)^2}{(\nu_C - 1)\left(\sum_{i=1}^{M} n_i^C V_{C1i}^{(t)}\right)^2}. \tag{8}$$

The hypothesis of testing for the overall treatment effect in the MRCT can thus be given as

$$\mathrm{H}_0 : \mu_T - \mu_C \leqq 0 \text{ vs. } \mathrm{H}_\mathrm{A} : \mu_T - \mu_C > 0$$

Although the hypothesis is one-sided, the method proposed can be straightforwardly extended to the two-sided hypothesis. Under the null hypothesis, the test statistic is given by

$$z = \frac{\tilde{\mu}_T - \tilde{\mu}_C}{\tilde{\sigma}_p}, \tag{9}$$

where $\tilde{\mu}_T$ and $\tilde{\mu}_C$ are estimated from the iterations of Eq. (6), $\tilde{\sigma}_p$ is from Eq. (8) and can be estimated by Eqs. (6) and (7). Under the null hypothesis, the test statistic follows a standard normal distribution for a large sample. Thus, we reject the null hypothesis at α level of significance if $z > z_{1\text{-}\alpha}$.

3 Sample Size Determination

Let N be the total sample size for each group, and p_i denote the proportion of patients out of $2N$ in the i^{th} region, $i = 1, \cdots, M$, where $\sum_{i=1}^{M} p_i = 1$. The total sample size for each group planned for detecting an expected treatment difference $\mu_T - \mu_C = \Delta$ at the desired significance level α and with power 1 - β will be obtained by solving

$$\frac{\Delta}{\tilde{\sigma}_p} - z_{1-\alpha} = z_{1-\beta}.$$

Assume that homogeneous variance condition holds. That is, $\sigma_{Ti} = \sigma_{Ci} = \sigma_i, \sigma_T = \sigma_C = \tilde{\sigma}$, $\nu_T = \nu_C = \tilde{\nu}$, $\lambda_T = \lambda_C = \tilde{\lambda}$, and thus we can assume that $V_{T1i} = V_{C1i} = V_{1i}$. Also assume that $n_i^T = n_i^C$, $i = 1, \ldots, M$. Let

$$\tilde{\sigma}_e = \sqrt{\frac{N}{2}}\tilde{\sigma}_p = \sqrt{\frac{N\tilde{\lambda}\tilde{\sigma}^2\tilde{\nu}\sum_{i=1}^{M} n_i \left(V_{1i}^{(t)}\right)^2}{(\tilde{\nu} - 1)\left(\sum_{i=1}^{M} n_i V_{1i}^{(t)}\right)^2}}.$$

Hence

$$\frac{2}{N}(Z_{1-\alpha}+Z_{1-\beta})^2\tilde{\sigma}_e^2=\Delta^2.$$

Consequently, we have

$$N=\frac{2(Z_{1-\alpha}+Z_{1-\beta})^2}{\Delta^2}\left(\frac{\tilde{\lambda}\tilde{\sigma}^2\tilde{\nu}\sum_{i=1}^{M}p_i(V_{1i}^{(t)})^2}{(\tilde{\nu}-1)(\sum_{i=1}^{M}p_iV_{1i}^{(t)})^2}\right). \tag{10}$$

Here, $V_{1i}^{(t)}$ can be estimated by the latent variable $\tau_i=\frac{\tilde{\sigma}}{\sigma_i}$ if we have sufficient information of σ_i^2 and $\tilde{\sigma}$ from early stage of development. Given $\tilde{\nu}$, $\tilde{\lambda}$, and Δ at the design stage, N can thus be derived by solving Eq. (10). However, in drug development, early stage trials are smaller and often limited to fewer regions. Even if not, the sample size from each region may be too small to get variance estimations with any precision. This may limit the application of our method.

4 Numerical Examples

To illustrate our approach, suppose an MRCT will be conducted in three regions (for example, Japan, EU, and the USA). Similar to Kawai et al. (2008), let p_1 be the proportion of patients in the smallest region, p_2 the proportion of patients in the second smallest region, and p_3 the proportion of patients in the largest region. Here, we consider three designs: $p_1=p_2<p_3$, $p_1<p_2=p_3$, and $p_1<p_2<p_3$.

By considering α=0.025, β=0.2, $\Delta=7$, $\tilde{\nu}=4$, $\tilde{\lambda}=1$, and $\tilde{\sigma}=20$, Tables 1 and 2 exhibit the total sample size and the assurance probability for different combinations of design parameters with $(\tau_1,\tau_2,\tau_3)=(0.7,0.9,1.4)$ and $(\tau_1,\tau_2,\tau_3)=(1.5,\ 1.1,\ 0.5)$, respectively. The first panel in the tables corresponds to $p_1=p_2<p_3$, the second panel corresponds to $p_1<p_2=p_3$, and the third panel corresponds to $p_1<p_2<p_3$. For instances, the first line in Table 1 corresponds to a design with (p_1, p_2, p_3)=(0.1, 0.1, 0.8), and $(\tau_1,\tau_2,\tau_3)=(0.7,\ 0.9,\ 1.4)$. In this case, the total sample size required per group in the MRCT is 204.

If $(\tau_1,\tau_2,\tau_3)=(0.7,\ 0.9,\ 1.4)$, the required sample size per group for the MRCT increases as p_1 increases. This makes intuitive sense, since for this case, Region 1 will have the largest variance. As p_1 increases, the overall variation $\tilde{\sigma}_e^2$ becomes larger. On the contrary, when $(\tau_1,\tau_2,\tau_3)=(1.5,\ 1.1,\ 0.5)$, the required sample size per group for the MRCT decreases as p_1 increases. In all cases, the total sample size required per group, N, is minimized when all regions are close to be equally represented. On the other hand, the shift in variation across regions we assume in Table 2 is larger than that in Table 1. Consequently, the difference between the sample sizes for all the scenarios in Table 2 (vary from 266 to 287) is larger than that in Table 1 (vary from 204 to 213).

Table 1 The required total sample size given $\alpha = 0.025$, $\beta = 0.1$, $\delta = 7$, $\sigma = 20$, $v = 8$, $\lambda = 1$ and $(\tau_1, \tau_2, \tau_3) = (0.7,\ 0.9,\ 1.4)$

$p_1 = p_2 < p_3$			
p_1	p_2	p_3	N
0.1	0.1	0.8	204
0.15	0.15	0.7	207
0.2	0.2	0.6	210
0.25	0.25	0.5	212
0.3	0.3	0.4	213
$p_1 < p_2 = p_3$			
p_1	p_2	p_3	N
0.1	0.45	0.45	209
0.15	0.425	0.425	210
0.2	0.4	0.4	211
0.25	0.375	0.375	212
0.3	0.35	0.35	213
$p_1 < p_2 < p_3$			
p_1	p_2	p_3	N
0.1	0.25	0.65	206
0.15	0.3	0.55	209
0.2	0.35	0.45	211

Table 2 The required total sample size given α=0.025, β=0.1, $\delta = 7$, $\sigma = 20$, $v = 4$, $\lambda = 1$ and $(\tau_1, \tau_2, \tau_3) = (1.5,\ 1.1,\ 0.5)$

$p_1 = p_2 < p_3$			
p_1	p_2	p_3	N
0.1	0.1	0.8	287
0.15	0.15	0.7	290
0.2	0.2	0.6	287
0.25	0.25	0.5	280
0.3	0.3	0.4	272
$p_1 < p_2 = p_3$			
p_1	p_2	p_3	N
0.1	0.45	0.45	267
0.15	0.425	0.425	268
0.2	0.4	0.4	268
0.25	0.375	0.375	268
0.3	0.35	0.35	266
$p_1 < p_2 < p_3$			
p_1	p_2	p_3	N
0.1	0.25	0.65	281
0.15	0.3	0.55	279
0.2	0.35	0.45	281

5 Discussions

Since the design of an MRCT incorporates subjects from many countries around the world, it might be reasonable to anticipate a difference in population variability due to ethnic differences. With consideration of the effect of ethnic differences on response variable in an MRCT, in this paper, we concern the case that the mean of response variable is fixed but the shift in the variance across regions is random. Our approach is different from other approaches assuming a common population standard deviation across regions. However, care needs to be taken when the shift in variation across regions is large. A larger shift in variation implies a larger regional difference. That is, the effect sizes among regions are very different. In this case, we need to consider whether or not the overall effect size in $(\tilde{\mu}_T - \tilde{\mu}_C)/\sigma_p$ is meaningful or interpretable for all regions. Therefore, if regional differences are suspected, we may only choose regions that likely meet the minimal clinical meaningfulness requirement of effect size. On the other hand, if regional differences are caused mostly by trial conduct/data quality among regions, the results of the entire trial might not be interpretable. Therefore, the statistical analysis plan should discuss specific elements of global trial needing considerations in design and analysis.

Our research work here is based on the assumption that the mean of response variable is fixed but the shift in the variance across regions is random. Chen et al. (2012) used a random effect model to model heterogeneous treatment effect across regions for the design and evaluation of an MRCT. For the most practical case of varying treatment effects with varying variances across regions, the normal-scaled inverse gamma distribution may be considered for modeling the unknown parameters. Future work is being pursued to address this issue.

References

Chen CT, Hung HMJ, Hsiao CF (2012) Design and evaluation of multi-regional trials with heterogeneous treatment effect across regions. Journal of Biopharmaceutical Statistic 22(5): 1037-1050

Edward RR, Doleys DM, Fillingim RB, Lowery D (2001) Ethnic differences in pain tolerance: clinical implications in a chronic pain population. Psychosomatic Medicine 63:316–323

Hung HMJ, Wang SJ, O'Neill R (2010) Consideration of regional difference in design and analysis of multi-regional trials. Pharmaceutical Statistics 9(3):173-178

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

Ko FS, Tsou HH, Liu JP, Hsiao CF (2010) Sample size determination for a specific region in a multi-regional trial. Journal of Biopharmaceutical Statistics 20:870-885

Lu Y, Chow SC, Zhang ZZ (2010) Statistical inference for clinical trials with random shift in scale parameter of target patient population. Duke Biostatistics and Bioinformatics (B&B) Working Paper Series: paper 11, Duke University, Durham, North Carolina, USA. Available at: http://biostats.bepress.com/dukebiostat/art11/ Accessed 17 December 2012

Tan CE, Emmanuel SC, Tan BY, Tai ES, Chew SK (2001) Diabetes mellitus abolishes ethnic differences in cardiovascular risk factors: lessons from a multi-ethnic population. Atherosclerosis 155:179–186

Part IX
In Vitro Drug Combination Studies

Experimental Design for In Vitro Drug Combination Studies

Gregory Hather, Huaihou Chen, and Ray Liu

Abstract In vitro drug combination studies typically involve a large number of wells with various concentrations of two drugs added together. To gain the most information from an experiment, what should the drug concentrations be? Here, we consider the case where the single drug response curves are known beforehand, but no previous data is available from the combination. We consider several designs, including C- and D-optimal designs and a factorial design. We evaluate these designs based on the expected variance of the synergy score for a large set of in vitro experiments performed at Takeda Pharmaceuticals. Based on the results, we were able to identify which design was the most efficient and robust.

1 Background

Drug combinations have become an important part of cancer care and antiviral therapy. To identify synergistic drug combinations, and to understand the combined behavior, scientists often perform in vitro drug combination studies. In the oncology setting, in vitro studies usually involve a cell viability assay applied to a cancer cell line. The assay usually involves a microtiter plate, where cells and various amounts of drugs are added to each well. The plate is then incubated, after which the cell viability is measured.

In the case where two drugs are considered, various methods have been used to analyze the data [1–4]. Some methods involve fitting a response surface model to

G. Hather (✉) • R. Liu
Takeda Pharmaceuticals, 35 Landsdowne Street, Cambridge, MA 02139, USA
e-mail: Greg.Hather@takeda.com; Ray.Liu@takeda.com

H. Chen
NewYork University, 1 Park Avenue, New York, NY 10016, USA
e-mail: hc2356@columbia.edu

M. Hu et al. (eds.), *Topics in Applied Statistics: 2012 Symposium of the International Chinese Statistical Association*, Springer Proceedings in Mathematics & Statistics 55,
DOI 10.1007/978-1-4614-7846-1_27,

describe the viability as a nonlinear function of the two drug concentrations [5, 6]. The fitted response surface can then be summarized by a single number to describe the synergy.

The choice of drug concentrations used in the experiment can affect the quality of the results. If the drug concentrations don't cover a reasonable range, then the response surface will be poorly estimated, so the synergy measure will be highly uncertain. In this paper, we will propose and evaluate several different designs for drug combination studies.

Experimental design for response surface estimation has a rich history in the literature. In particular, design methods for nonlinear response surface surfaces have been explored, both in a general context [7, 8] and in the context of drug combination studies [9, 10]. In the nonlinear setting, the optimal design for finding the response surface parameters actually depends on the parameters. If a small-scale experiment has been done previously, then the parameter estimates from this experiment can be used to design a larger experiment. However, combination studies are often performed without previous combination data. In this paper, we consider the case where previous combination data is unavailable, but where previous single agent data has been collected.

2 Analyzing Cell Viability Data

Takeda scientists performed a number of combination studies using 384-well microtiter plates (Fig. 1). To analyze this data, we normalized the viability by scaling so that the median of the negative controls was 0 and the median of the positive controls was 100. More formally,

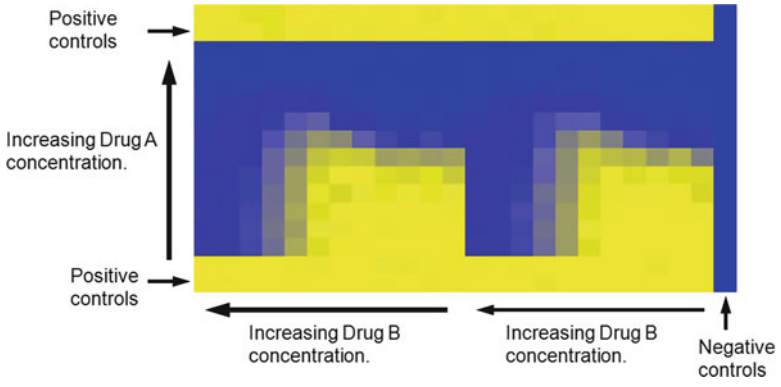


Fig. 1 A heatmap showing the viability measurements for a drug combination study performed at Takeda Pharmaceuticals. Here, *yellow* corresponds to high viability and *blue* corresponds to low viability

$$V_i = 100\frac{U_i - \text{median}(U_-)}{\text{median}(U_+) - \text{median}(U_-)},$$

where V_i is the normalized viability of the ith well, and U_i is the corresponding raw viability measurement. After normalization, the controls were discarded.

2.1 Single Drug Experiments

Some of the experiments only involved single agents. For these cases, we assumed that the dose response curve had the form of the Hill equation [11]. Since the data was normalized, we assumed that the highest viability (the upper plateau) was 100. Thus, the viability was modeled as

$$V = 100 - \frac{E_{\max}}{1 + (I/C)^S} + \text{error},$$

where V is the normalized viability measurement. Here, $E_{\max}$ is the maximum drug effect, I is the inflection point, S is the slope, and C is the drug concentration. We assumed that the error values were identically distributed normal random variables that were independent of each other. These assumptions can be tested by exploratory data analysis and residual analysis. We used the nlm() function in the R software package [12] to minimize the sum of the squared residuals and estimate the lower plateau, the slope, and the inflection point.

2.2 Combination Experiments

To describe the relationship between the normalized viability and the drug concentrations, we used a response surface model similar to that of [5], which is an extension of the Hill equation. For a given plate, let

$$\begin{aligned} C &= (C_A/I_1) + (C_B/I_2) \\ x &= (C_A/I_1)/C \\ E_{\max} &= E_1 + E_2 x + E_3 x^2 + E_4 x^3 \\ I &= 1 + I_3 x(1-x) \\ S &= S_1 + S_2 x + S_3 x^2 + S_4 x^3 \\ V &= 100 - \frac{E_{\max}}{1 + (I/C)^S} + \text{error}, \end{aligned}$$

where E_1, E_2, E_3, E_4, I_1, I_2, I_3, S_1, S_2, S_3, and S_4 are parameters, C_A and C_B are the respective concentrations of drugs A and B, and V is the normalized viability measurement. This model has the property that along any line of constant dose ratio, the model has the form of a Hill equation. We assumed that the error values were independent and identically distributed normal random variables. We also included two constraints, thus yielding a model with 9 degrees of freedom.

We fit the data to our model by minimizing the residual sum of squares using the used the nlm() function in R. Next, we developed a measure of synergy that was a function of the nine response surface parameters. We refer to this measure as the synergy score.

3 Designing Combination Experiments

For a given number of wells, we wish to choose doses that will minimize the variance of the synergy measure. Unfortunately, the best choice depends on the shape of the response surface, which we don't have until we do the experiment. However, if we have past single drug data, and we assume that there is no interaction, then we have a guess for the parameters. For a model with no interaction, we assume that $I_3 = 0$. We also assume that for $x = 0.5$, the resulting slope and E_{max} are found by averaging the slope and E_{max}, respectively, for the individual drugs. With these constraints, one can uniquely identify parameters for the model.

3.1 A C-Optimal Design

Figure 2 shows a strategy for finding an optimal study design. Given single agent data, one can produce a guess for the response surface parameters. In addition, given the dose choices for a proposed design, one can compute the likelihood function and the Fisher Information matrix, evaluated at the initial parameter guess. Using the Cramer-Rao lower bound [13, 14], the Fisher Information matrix can be used to estimate the variance matrix of the parameter estimates with the proposed design. Since the synergy score is simply a function of the response surface parameters, the one can use the Delta method [13] to estimate the expected variance of the synergy score under the proposed design. Then, one can adjust the proposed design to minimize the estimated variance.

The Delta method uses a first order (i.e., linear) approximation to relate the synergy score to the response surface parameters. Therefore, this approach minimizes the variance of a linear combination of the parameters. In the literature, this is called a C-optimal design [7].

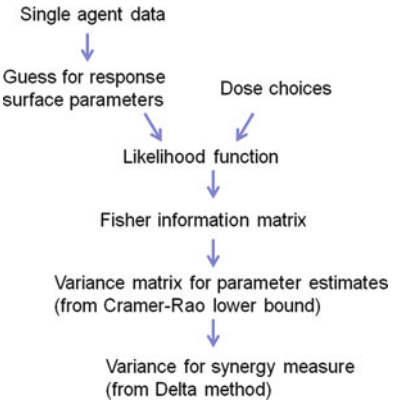


Fig. 2 The process for choosing a design

3.2 Computational Approach

The experimental design specifies the concentration for each of the two drugs in each microtiter well. Thus, minimizing the estimated variance of the synergy score would require a search over a space with dimension equal to twice the number of wells. Finding the global optimum in this space would require far too much computational power. Therefore, we developed a heuristic approach to this optimization problem.

Start with 6×6 log-spread grid points, search over the 50×50 candidate grid points, and choose the point which minimizes the variance of the synergy measure estimate. Repeat the above search step until we obtain the number of design points needed. Remove the 36 starting grid points and search another 36 design points over the 50×50 candidate grid points.

3.3 Evaluating Designs

To evaluate our C-optimal design procedure, we obtained data from a set of 100 drug combination studies performed at Takeda Phamaceuticals using a variety of drugs and cancer cell lines. The studies were done with 384 well microtiter plates. The doses were arranged in a factorial design, with dose ranges manually chosen by the scientists. In addition, we obtained single agent data for each drug from previous experiments.

Figure 3 shows our method for evaluating the design. For each of the 100 experiments, we found the C-optimal design based on the past single agent data. Next, we took the corresponding combination data, estimated the response surface parameters, and used these estimates to predict the variance of the synergy measure. We believe this is a realistic assessment of the design procedure because it uses real

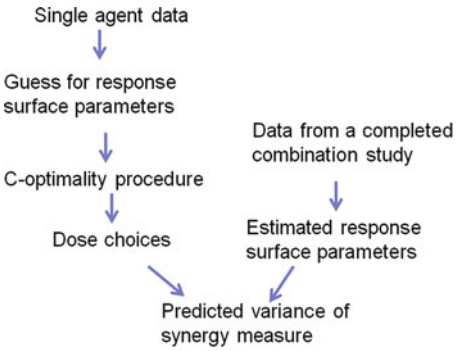


Fig. 3 The process for evaluating the C-optimal design

single agent data to create the designs and real combination data to evaluate the designs.

3.4 Other Designs

In addition to the C-optimal design, we considered two other designs, which we call the D-optimal design and the automated factorial design. The D-optimal design [7] is similar to the C-optimal design, except it minimizes the determinate of the estimated parameter covariance matrix.

The automated factorial design is a factorial design with dose ranges chosen based on the past single agent data. From the past single agent data, we estimate the slope and inflection point, and we used the following formula to select a dose range.

$$\text{Log dose range} = \left(\log(I) - a - \frac{b}{\text{slope}}, \log(I) + a + \frac{c}{\text{slope}}\right),$$

where I is the inflection point, and a, b, and c are constants. The dose levels are evenly distributed on the log scale along the selected dose range. Note that the dose range is centered around the inflection point, and the width of the range decreases with increasing slope. This ensures that the dose range covers the region when the response is changing.

4 Results and Discussion

The results are shown in Table 1. We found that the C-optimal design actually underperformed the designs manually created by the scientists. The D-optimal design is only slightly better than the manual designs. We believe these optimal designs underperform because there is a difference between the parameters used to

Table 1 Predicted variance of the estimate for the synergy score for the various designs

Experiment	C-optimal	D-optimal	Automated factorial
1	0.63	1.00	0.43
2	1.02	0.38	0.66
3	1.26	0.25	0.56
4	1.31	0.52	0.72
5	2.04	0.06	0.26
6	2.18	0.60	0.45
7	0.29	0.55	0.62
8	0.19	0.94	0.17
9	2.64	0.73	0.33
100	0.98	0.45	0.45
Mean	4.60	0.83	0.49

The variance is expressed as a fraction of the variance expected under the manual factorial designs

Table 2 Predicted variance of the estimate for the synergy score for the C- and D-optimal designs

Experiment	C-optimal	D-optimal
1	0.10	0.22
2	0.12	0.24
3	0.08	0.20
4	0.11	0.34
5	0.04	0.10
6	0.05	0.15
7	0.15	0.39
8	0.02	0.004
9	0.02	0.07
100	0.21	0.45
Mean	0.112	0.216

In this case, the designs were evaluated using the same parameters that were used to generate the designs. The variance is expressed as a fraction of the variance under the manual factorial designs

generate the designs and the parameters used to evaluate the designs. The parameters used to generate the designs are found using the prior single agent data. This data is from a different batch than the combination data, so the parameter values may have shifted. Also, the initial parameter guesses assume that there is no drug synergy. Therefore, the initial parameter guesses may differ substantially from the parameters estimated from the combination data. We believe the C- and D-optimal designs are not robust to these differences. Furthermore, previous studies have found that C- and D-optimal designs can perform poorly if there is a high level of uncertainty in the initial parameter guesses [8].

To confirm our view, we evaluated the C- and D-optimal designs using the same parameters that were used to generate these designs. The results are shown in Table 2. As expected, the C- and D-optimal designs perform well in this scenario.

In Table 1, the automated factorial design outperforms the manual factorial designs. The variance of the synergy measure is reduced by a factor of 2, which means the number of wells could be cut in half if the automated factorial design is used.

5 Conclusion

We presented a method to evaluate different designs for drug combination studies. We found that the C-and D-optimal designs were not robust to misspecification of the response surface parameters. The automated factorial design showed a twofold reduction in the variance of the synergy measure.

Acknowledgments We would like to thank the Double Agent team at Takeda Pharmaceuticals for their support.

References

[1] Chou, T.C., Talalay, P.: Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors. Adv. Enzyme. Reg. **22**, 27-55 (1984)
[2] Berenbaum, M.C.: The expected effect of a combination of agents: The general solution. J. Theor. Biol. **114**, 413-431 (1985)
[3] Peterson, J.J., Novik, S.J.: Nonlinear Blending: A Useful General Concept for the Assessment of Combination Drug Synergy. J. Recept. Signal. Transduct. Res. **27**, 125-146 (2007)
[4] Bliss, C.I.: The toxicity of poisons combined jointly. Ann. Appl. Biol. **26**, 585-615 (1939)
[5] Minto, C.F., Schnider, T.W., Short, T.G., Gregg, K.M., Gentilini, A., Shafer, S.L.: Response Surface Model for Anesthetic Drug Interactions. Anesthesiology **92**, 1603-1616 (2000).
[6] Greco, W.R., Park, H.S., Rustum, Y.M.: Application of a New Approach for the Quantitation of Drug Synergism to the Combination of cis-Diamminedichloroplatinum and 1-b-d-Arabinofuranosylcytosine. Cancer Res **50**, 5318-5327 (1990)
[7] Pukelsheim, F.: Optimal design of experiments. New York: J. Wiley. (1993).
[8] Chaloner, K., and Verdinelli, I.: Bayesian Experimental Design: A Review. Statistical Science **10**, 273-304 (1995)
[9] Tallarida, R.J., Stone, D.J. Jr., Raffa. R.B.: Efficient designs for studying synergistic drug combinations. Life Sciences **61**, PL417-425 (1997)
[10] Chou, T.C.: Theoretical Basis, Experimental Design, and Computerized Simulation of Synergism and Antagonism in Drug Combination Studies. Pharmacol. Rev. **58**, 621-681 (2006)
[11] Hill, A.V.: The possible effects of the aggregation of the molecules of hæmoglobin on its dissociation curves. J. Physiol. **40**, iv-vii (1910)
[12] R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. (2008)
[13] Cramer, H.: Mathematical Methods of Statistics. Princeton, NJ: Princeton Univ. Press. (1946)
[14] Rao, C.R. Information and the accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society **37**, 81-89 (1945)