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

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

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

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

Springer

ICSA Book Series in Statistics

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

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

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

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

Springer

Editors
Jianchang Lin
Global Statistics
Takeda Pharmaceuticals
Cambridge, MA, USA

Bushi Wang
Biostatistics
Boehringer Ingelheim Pharmaceuticals
Ridgefield, CT, USA

Xiaowen Hu
Colorado State University
Fort Collins, CO, USA

Kun Chen
University of Connecticut
Storrs, CT, USA

Ray Liu
Global Statistics
Takeda Pharmaceuticals
Cambridge, MA, USA

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

Library of Congress Control Number: 2016954077

Printed on acid-free paper

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

Preface

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

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

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

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

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

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

Contents

Contributors

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Jianchang Lin Takeda Pharmaceuticals, Cambridge, MA, USA

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Serap Sankoh Takeda Pharmaceuticals, Cambridge, MA, USA

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

Leslie Sidor Biogen, Cambridge, MA, USA

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

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

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

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

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

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

Zhaoyang Teng Takeda Pharmaceuticals, Cambridge, MA, USA

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Bin Zhang Genocea Biosciences, Cambridge, MA, USA

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

Yuanye Zhang Agios Pharmaceuticals, Cambridge, MA, USA

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

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

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

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

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

Part I
Biomarker and Personalized Medicine

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

Yong Zang and Ying Yuan

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

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

1 Introduction

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

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

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

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

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

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

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

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

2 Methods

2.1 Optimal Allocation Rules

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

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

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

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

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

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

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

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

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

2.2 Likelihood Ratio Test Based on EM Algorithm

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

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

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

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

At the M-step, we update

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

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

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

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

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

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

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

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

3 Simulation Studies

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

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

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

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

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

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

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

All values are in percentages

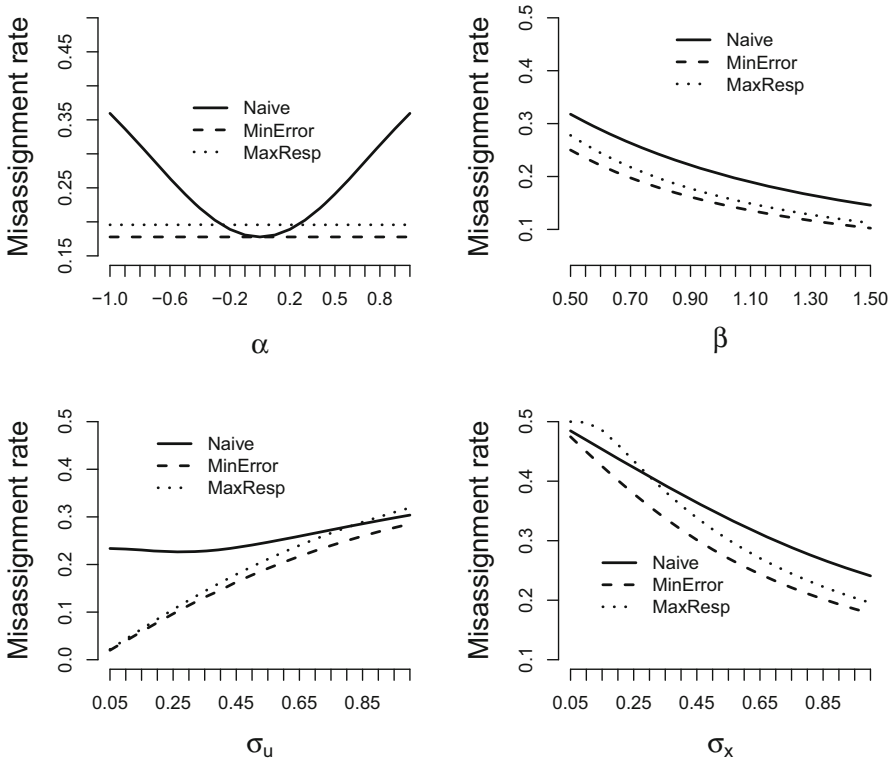

Fig. 1 The misassignment rates of the naive, MinError and MaxResp designs

4 Conclusion

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

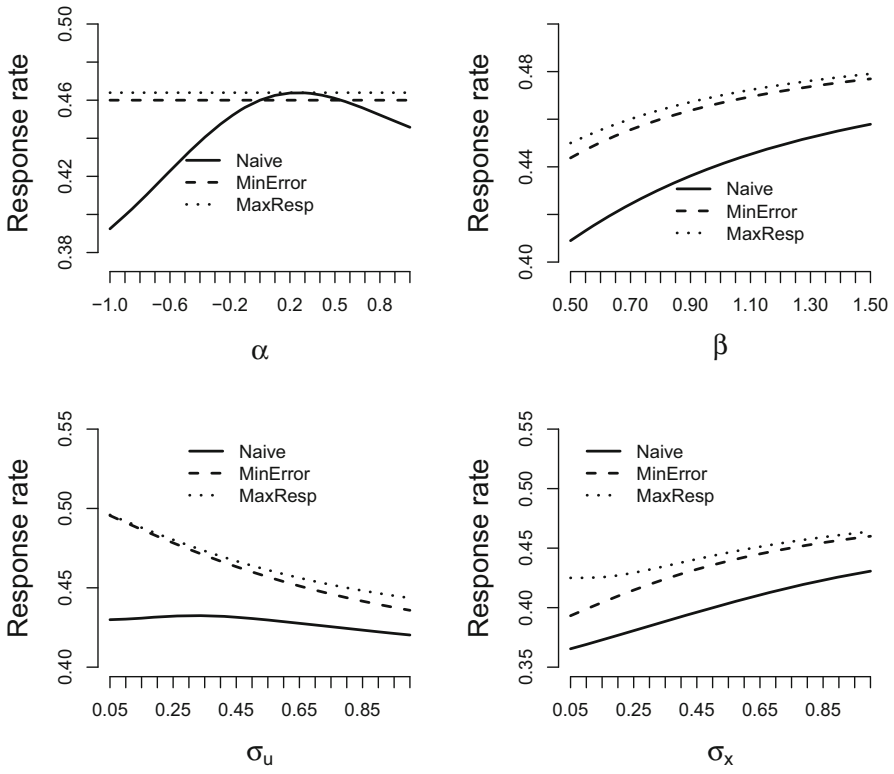

Fig. 2 The response rates of the naive, MinError and MaxResp designs when $p_{10} = p_{01} = 0.1$ and $p_{00} = 0.2$

Appendix

Proof of Theorem 1

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

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

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

Proof of Theorem 2

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

$$
\begin{aligned}
&\text{pr}(Y = 1) = \sum_{j=0}^{1}\sum_{k=0}^{1}\text{pr}(M = j, T = k)p_{jk} \\
&= C + D_0\int \text{pr}(M = 0|W)\text{pr}(T = 0|W)f(W)dW \\
&\quad + D_1\int \text{pr}(M = 1|W)\text{pr}(T = 1|W)f(W)dW \\
&= C + D_0\int (1 - \text{pr}(M = 1|W))(1 - \text{pr}(T = 1|W))f(W)dW \\
&\quad + D_1\int \text{pr}(M = 1|W)\text{pr}(T = 1|W)f(W)dW \\
&= C + \int [D_0\{1 - \text{pr}(M = 1|W)\} + \text{pr}(T = 1|W)\{D_1 - \omega\text{pr}(M = 0|W)\}]f(W)dW.
\end{aligned}
$$

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

References

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

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

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

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

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

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

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

Statistical Considerations for Evaluating Prognostic Biomarkers: Choosing Optimal Threshold

Zheng Zhang

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

Keywords Biomarker • ROC • Threshold • Optimal • Youdan index

1 Introduction

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

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

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

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

2 A Brief Review of the ROC Curve

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

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

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

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

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

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

3 Criteria Based on the ROC Curve

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

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

4 Issues When Reporting the Optimal Threshold

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

5 Simulation Study

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

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

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

Table 1 Thresholds and the associated accuracy measures

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

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

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

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

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