Journal of Probability and Statistics

New Advances in Biostatistics

Lead Guest Editor: Yichuan Zhao
Guest Editors: Ash Abebe, Lihong Qi, Min Zhang, and Xu Zhang

New Advances in Biostatistics

Journal of Probability and Statistics

New Advances in Biostatistics

Lead Guest Editor: Yichuan Zhao
Guest Editors: Ash Abebe, Lihong Qi, Min Zhang, and Xu Zhang

Editorial Board

Contents

Hindawi
Journal of Probability and Statistics
Volume 2019, Article ID 1352310, 2 pages
https://doi.org/10.1155/2019/1352310

Editorial

New Advances in Biostatistics

Yichuan Zhao,[1] Ash Abebe,[2] Lihong Qi,[3] Min Zhang,[4] and Xu Zhang[5]

[1] *Department of Mathematics and Statistics, Georgia State University, Atlanta, GA, USA*
[2] *Department of Mathematics and Statistics, Auburn University, Auburn, AL, USA*
[3] *Division of Biostatistics, Department of Public Health Sciences, University of California, California, Davis, CA, USA*
[4] *Department of Statistics, Purdue University, West Lafayette, IN, USA*
[5] *Department of Internal Medicine, Medical School, University of Texas Health Science Center at Houston, Houston, TX, USA*

Correspondence should be addressed to Yichuan Zhao; yichuan@gsu.edu

Received 2 May 2019; Accepted 5 May 2019; Published 16 May 2019

Biostatistics deals with data arising from biomedical research. It remains a very active research area with complicated time-to-event data and missing data emerging in application areas including medicine, genetics, neuroscience, and engineering. Recent advances in biomedical research have created new challenges and opportunities for statisticians and data scientists. For example, big data analysis, precision medicine, artificial intelligence, causal inference, and other new research fields have inspired data scientists to develop modern statistical methods and innovative inference procedures.

The present special issue presents statistical research and new advances in contemporary biostatistics. It provides one review article and seven research articles contributed by some of the leading experts in the field. The review article contained in this issue gives an overview of the current development in biostatistics, while each of the seven research articles addresses new challenges in contemporary biostatistics, as summarized below.

In the review paper "Mixed Effects Models with Censored Covariates, with Applications in HIV/AIDS Studies," by L. Wu and H. Zhang, the authors focus on the problem of censored time-dependent covariates in regression analysis of longitudinal data and time-to-event data. They review the two-step method and the joint likelihood method and describe examples from HIV/AIDS studies to illustrate the problem and applications of the methods.

The research article entitled "A Comparison of Mean-Based and Quantile Regression Methods for Analyzing Self-Report Dietary Intake Data," by M. L. Vidoni et al., compares the traditional mean-based linear regression with quantile regression in terms of investigating the relationship between health behavior intervention and eating indices. The authors found that only the quantile regression, through modeling the coefficients across distributions of the outcome, can fully describe the effect of intervention on healthy and unhealthy eating indices between intervention and standard care groups. The results can help develop more effective behavioral intervention trails with heterogeneous populations.

The research article entitled "Atrial Fibrillation Detection by the Combination of Recurrence Complex Network and Convolution Neural Network," by X. Wei et al., proposes an R wave peak interval independent atrial fibrillation detection algorithm on the basis of analyzing the synchronization features of the electrocardiogram signal by a deep neural network. Results show that the sensitivity, specificity, and accuracy of the algorithm are all around 95%, and the algorithm is more effective than the traditional algorithms in terms of detecting individual variation in the atrial fibrillation.

The research paper entitled "A Note on the Adaptive LASSO for Zero-Inflated Poisson Regression," by P. Banerjee et al., proposes a flexible variable selection approach to efficiently identify correlated features in a zero-inflated Poisson (ZIP) regression model. The existing approach for variable selection in a ZIP regression model which satisfies the oracle property is the EM adaptive LASSO (EM AL), which generates suboptimal results when the features are correlated. The proposed approach is able to identify the true model consistently, and the resulting estimator is as efficient as oracle.

The past decade has seen extensive development of statistical methodology for designing phase I clinical trials for drug combinations including designs allowing individualized maximum tolerable dose (MTD) determination in single agent cancer phase I trials. In the research article "A Bayesian Adaptive Design in Cancer Phase I Trials Using Dose Combinations in the Presence of a Baseline Covariate," by M. A. Diniz et al., the authors describe a Bayesian adaptive design for dose finding of a combination of two drugs in cancer phase I clinical trials. The method takes into account patients' heterogeneity possibly related to treatment susceptibility. The authors accomplish this using escalation with overdose control principle, and the proposed method gives a smaller pointwise average bias and a higher percent of MTD recommendation.

In the research paper "Improved Small Sample Inference on the Ratio of Two Coefficients of Variation of Two Independent Lognormal Distributions," by A. Wong and L. Jiang, the authors study the two-sample inference for the ratio of two coefficients of variation where the data is sampled from lognormal distributions. They propose a simulated likelihood ratio method that outperforms existing methods with small samples in simulation studies.

In the research article "Detecting Spatial Clusters via a Mixture of Dirichlet Processes," by M. A. Ray et al., the authors propose an approach able to detect spatial clusters with skewed or irregular distributions. A mixture of Dirichlet processes is used to describe spatial distribution patterns. The effects of different batches of data collection efforts are also modeled with a Dirichlet process. Inferences of parameters including clustering are drawn under a Bayesian framework.

The research article entitled "On the Use of Min-Max Combination of Biomarkers to Maximize the Partial Area under the ROC Curve," by H. Ma et al., adopts and extends the min-max method to the estimation of the pAUC when multiple continuous scaled biomarkers are available and compare the performances of the proposed approach with existing approaches via simulations. The extensive simulation results demonstrate that the proposed method provides the largest pAUC estimates. The proposed method is robust, and it is encouraged to use this approach in the estimation of the pAUC for many practical scenarios.

As the editors of this special issue, we hope that readers of this special issue will find these articles representative of the contributions of the contemporary biostatistics, in terms of statistical procedures and practical applications. We hope that the special issue provides new methods and novel applications motivated by biomedical examples in the broad areas of biostatistics and stimulate new interests in contemporary biostatistics.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We would like to express our appreciations to the authors and reviewers for their excellent contributions to this special issue. We also thank the Journal of Probability and Statistics for the professional service, which makes this special issue possible.

Yichuan Zhao
Ash Abebe
Lihong Qi
Min Zhang
Xu Zhang

Hindawi
Journal of Probability and Statistics
Volume 2019, Article ID 7173416, 7 pages
https://doi.org/10.1155/2019/7173416

Research Article

Improved Small Sample Inference on the Ratio of Two Coefficients of Variation of Two Independent Lognormal Distributions

A. Wong [1] and L. Jiang [2]

[1] *Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3*
[2] *Beijing Education Examinations Authority, Beijing, China*

Correspondence should be addressed to A. Wong; august@yorku.ca

Received 5 July 2018; Revised 29 October 2018; Accepted 27 January 2019; Published 3 March 2019

Guest Editor: Xu Zhang

Without the ability to use research tools and procedures that yield consistent measurements, researchers would be unable to draw conclusions, formulate theories, or make claims about generalizability of their results. In statistics, the coefficient of variation is commonly used as the index of reliability of measurements. Thus, comparing coefficients of variation is of special interest. Moreover, the lognormal distribution has been frequently used for modeling data from many fields such as health and medical research. In this paper, we proposed a simulated Bartlett corrected likelihood ratio approach to obtain inference concerning the ratio of two coefficients of variation for lognormal distribution. Simulation studies show that the proposed method is extremely accurate even when the sample size is small.

1. Introduction

In health and medical research, it is common that the variable of interest, X, such as the survival time, takes only positive values and the underlying distribution of this variable is highly skewed to the right. In this case, the frequently assumed normal distribution for X is not suitable. A standard approach to first transform X such that the transformed variable $Y = g(X)$ is normally distributed. Then the existing statistical theories developed for the normal distribution can be applied. For $X > 0$ and the distribution of X is highly skewed to the right, the most common transformation is the logarithmic transformation. In other words, $Y = \log(X)$ is normally distributed. Hence, X is lognormally distributed. Detailed review of the theories of the lognormal distribution can be found in Aitchison and Brown [1], and Crow and Simizu [2]. In practice, Fears et al. [3] investigated the variability and reproducibility of hormone assays used by laboratories with the capability of performing large numbers of tests. They assumed the hormone samples used in laboratories are independent lognormally distributed. In this case, it is of special interest to know if each sample yields consistent measurements.

The coefficient of variation (τ) is defined as the ratio of the standard deviation to the mean, where the mean is assumed to be non zero. It is an important index for assessment of the reliability of a measuring procedure. Hence, the problem considered in Fears et al. [3] can be viewed as testing if the coefficients of variation used in each laboratory are the same or not.

Mathematically, if a random variable X is distributed as lognormal(μ, σ), then $Y = \log(X)$ is distributed as normal with mean μ and variance σ^2. It is well-known that

$$E(X) = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}, \quad \text{and } \operatorname{var}(X) = \exp\left\{2\mu + \sigma^2\right\}\left[\exp\left\{\sigma^2\right\} - 1\right]. \tag{1}$$

Hence, the coefficient of variation, τ, is

$$\tau = \frac{\sqrt{\operatorname{var}(X)}}{E(X)} = \sqrt{\exp\left\{\sigma^2\right\} - 1} > 0. \tag{2}$$

Nam and Kwon [4] compared various approximate interval estimations of the ratio of two coefficients of variation for independent lognormal distributions. And their simulation results showed that empirical coverage rates of these methods are satisfactorily close to the nominal coverage rate for medium sample size. The aim of this paper is to develop a more accurate method to obtain inference for the ratio of two coefficients of variation for independent lognormal distributions. Moreover, the proposed method can be generalized to test if the coefficients of variation from k independent lognormal distributions are heterogeneous.

The rest of the paper is organized as follows. Section 2 reviewed the existing methods for obtaining inference concerning the ratio of two coefficients of variation from independent lognormal distribution. The simulated Bartlett corrected likelihood method is proposed in Section 3. A real data example is presented in Section 4 to illustrate the application of the method discussed in this paper. Simulation studies are performed to compare the accuracy of the methods discussed in this paper in Section 5. Extension to testing for homogeneity of coefficients of variations from k independent lognormal distributions is discussed in Section 6. Some concluding remarks are recorded in Section 7.

2. Existing Methods for Inference on the Ratio of Two Coefficients of Variation of Two Independent Lognormal Distributions

Let $(x_{i1},\ldots,x_{in_i})$ be the i^{th} sample from the lognormal(μ_i,σ_i) distribution, where $i = 1,\ldots,k$. Then $(y_{i1},\ldots,y_{in_i}) = (\log x_{i1},\ldots,\log x_{in_i})$ is the i^{th} sample from the normal distribution with mean μ_i and variance σ_i^2. From (2), the i^{th} coefficient of variation is $\tau_i = \sqrt{\exp\{\sigma_i^2\}-1}$. Nam and Kwon [4] compared four methods in obtaining confidence intervals for $\psi = \tau_1/\tau_2$. The following is the summary of the methods discussed in Nam and Kwon [4]:

(1) Wald type method

Let the observed test statistic be

$$z_W(\psi) = \frac{\widehat{\psi}-\psi}{\sqrt{\widehat{\operatorname{var}}(\widehat{\psi})}} \tag{3}$$

where $\widehat{\psi} = \widehat{\tau}_1/\widehat{\tau}_2$, $\widehat{\tau}_i = \sqrt{\exp\{\widehat{\sigma}_i^2\}-1}$, $\widehat{\sigma}_i^2 = (1/n_i)\sum_{j=1}^{n_i}(\log x_{ij} - \sum_{h=1}^{n_i}\log x_{ih}/n_i)^2$, and $\widehat{\operatorname{var}}(\widehat{\psi}) = (n_1\psi^4[(1+\widehat{\tau}_2^2)\widehat{\sigma}_2^2]^2 + n_2[(1+\widehat{\tau}_1^2)\widehat{\sigma}_1^2]^2)/2n_1n_2(\widehat{\tau}_1\widehat{\tau}_2)^2$. Then $Z_W(\psi)$ is asymptotically distributed as standard normal distribution. The significance function of ψ is $p(\psi) = \Phi(z_W(\psi))$, where $\Phi()$ is the cumulative distribution function of the standard normal distribution.

(2) Fieller type method

Let the observed test statistic be

$$z_F(\psi) = \frac{\widehat{\tau}_1-\psi\widehat{\tau}_2}{\sqrt{\widehat{\operatorname{var}}(\widehat{\tau}_1)+\psi^2\widehat{\operatorname{var}}(\widehat{\tau}_2)}} \tag{4}$$

where

$$\widehat{\operatorname{var}}(\widehat{\tau}_i) = \frac{\widehat{\sigma}_i^2(1+\widehat{\tau}_i)^2}{2n_i\widehat{\tau}_i^2} \quad i = 1, 2. \tag{5}$$

Then $Z_F(\psi)$ is also asymptotically distributed as standard normal distribution. The significance function of ψ is $p(\psi) = \Phi(z_F(\psi))$.

(3) Log method

Let the observed test statistic be

$$z_L(\psi) = \frac{\log\widehat{\tau}_1-\log\widehat{\tau}_2-\log\psi}{\sqrt{\widehat{\operatorname{var}}(\log\widehat{\tau}_1)+\widehat{\operatorname{var}}(\log\widehat{\tau}_2)}} \tag{6}$$

where

$$\widehat{\operatorname{var}}(\log\widehat{\tau}_{\mathrm{i}}) = \frac{\widehat{\operatorname{var}}(\widehat{\tau}_i)}{\widehat{\tau}_i^2} \quad i = 1, 2. \tag{7}$$

Then $Z_L(\psi)$ is also asymptotically distributed as standard normal distribution. The significance function of ψ is $p(\psi) = \Phi(z_L(\psi))$.

(4) Method of variance estimates recovery (MOVER)

This is a method that will directly obtain an approximate $(1-\alpha)100\%$ confidence interval for ψ only. Let

$$\begin{aligned} l_i &= \log\widehat{\tau}_i - z_{\alpha/2}\sqrt{\widehat{\operatorname{var}}(\log\widehat{\tau}_i)}, \\ \text{and } u_i &= \log\widehat{\tau}_i + z_{\alpha/2}\sqrt{\widehat{\operatorname{var}}(\log\widehat{\tau}_i)}. \end{aligned} \tag{8}$$

Then an approximate $(1-\alpha)100\%$ confidence interval for $\log\psi$ is (L,U) where

$$\begin{aligned} L &= (\log\widehat{\tau}_1 - \log\widehat{\tau}_2) \\ &\quad - \sqrt{(\log\widehat{\tau}_1 - l_1)^2 + (u_2 - \log\widehat{\tau}_2)^2}, \\ U &= (\log\widehat{\tau}_1 - \log\widehat{\tau}_2) \\ &\quad + \sqrt{(\log\widehat{\tau}_1 - l_1)^2 + (u_2 - \log\widehat{\tau}_2)^2}. \end{aligned} \tag{9}$$

Thus, an approximate $(1-\alpha)100\%$ confidence interval for ψ is $(\exp\{L\},\exp\{U\})$. If $\widehat{\operatorname{var}}(\log\widehat{\tau}_i)$, for $i = 1, 2$, to be the same as that obtained in the Log method, the MOVER method is identical to the Log method. Note that Hasan and Krishamoorthy [5] proposed an improved version of the MOVER method.

3. Proposed Method

In this section, we will first review the likelihood based methods and the Bartlett corrected likelihood ratio method. Since the required Bartlett adjustment for the Bartlett corrected likelihood ratio method is very difficult to obtain, a numerical algorithm is proposed to approximate the Bartlett adjustment. Then the methods are applied to obtain inference for the ratio of two coefficients of variaation of two independent lognormal distribution.

3.1. Likelihood Based Methods and Bartlett Corrected Likelihood Ratio Method. Let $(x_1, \ldots, x_n)$ be a sample from a known distribution with probability density function $f(\cdot, \theta)$, where θ is a p-dimensional vector of parameters. Let $\psi = \psi(\theta)$, which has dimension $d < p$ be the parameter of interest. The log-likelihood function is

$$\ell(\theta) = \ell(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i; \theta). \tag{10}$$

Under the regularity conditions stated in Barndorff-Nielsen and Cox [6], we have the standardized maximum likelihood estimate (MLE) statistic $(\widehat{\theta} - \theta)'[\text{var}(\widehat{\theta})]^{-1}(\widehat{\theta} - \theta)$ and the likelihood ratio statistic $2[\ell(\widehat{\theta}) - \ell(\theta)]$ that are asymptotically chi-square distributed with p degrees of afreedom, χ^2_p, where $\widehat{\theta}$ is the overall MLE, which is the value of θ that maximized $\ell(\theta)$, and $\text{var}(\widehat{\theta})$ is approximately the inverse of the Fisher's expected information. When the parameter of interest is $\psi = \psi(\theta)$, Barndorff-Nielsen and Cox [6] showed that similar statistics can be obtained. The standardized MLE statistic becomes

$$Q(\psi) = (\widehat{\psi} - \psi)' [\text{var}(\widehat{\psi})]^{-1} (\widehat{\psi} - \psi) \tag{11}$$

where $\widehat{\psi} = \psi(\widehat{\theta})$, and $\text{var}(\widehat{\psi})$ can be approximated by the delta method, which takes the form

$$\text{var}(\widehat{\psi}) \approx \left\{ \frac{\partial \psi(\theta)}{\partial \theta} \right\}' \text{var}(\widehat{\theta}) \left\{ \frac{\partial \psi(\theta)}{\partial \theta} \right\}. \tag{12}$$

The likelihood ratio statistic is

$$W(\psi) = 2\left[\ell(\widehat{\theta}) - \ell(\widetilde{\theta})\right], \tag{13}$$

where $\widetilde{\theta}$ is the constrained MLE, which is obtained by maximizing $\ell(\theta)$ for the given ψ value. Both $Q(\psi)$ and $W(\psi)$ are asymptotically χ^2_d. As defined in Fraser [7], the significance function for ψ is defined as $p(\psi) = P(\chi^2_d \le q(\psi))$ or $p(\psi) = P(\chi^2_d \le w(\psi))$ can be used to obtain inference concerning ψ where $q(\psi)$ and $w(\psi)$ are the observed values of $Q(\psi)$ and $W(\psi)$, respectively. In particular, the $(1-\alpha)100\%$ confidence region of ψ is

$$\begin{gathered} \left\{\psi : q(\psi) \le \chi^2_{d,1-\alpha}\right\} \\ \text{and } \left\{\psi : w(\psi) \le \chi^2_{d,1-\alpha}\right\}, \end{gathered} \tag{14}$$

respectively, where $\chi^2_{d,1-\alpha}$ is the $(1-\alpha)100^{th}$ percentile of χ^2_q.

It is well-known that these two asymptotic methods have rate of convergence $O(n^{-1/2})$, and they are referred to as the first-order methods. In statistics literature, there exists various adjustments to improve the accuracy of the above methods. In particular, Barndorff-Nielsen [8, 9] introduced the modified signed log-likelihood ratio statistics, a third-order method. However, this method is restricted to scalar parameter of interest only. On the other hand, Bartlett [10] proposed a transformation of the likelihood ratio statistic such that the mean of the transformed statistic matched the mean of the asymptotic distribution. More specifically,

$$W^*(\psi) = \frac{W(\psi)}{B} \tag{15}$$

where B is the Bartlett adjustment such that $E[W^*(\psi)] = d$. And $W^*(\cdot)$ is known as the Bartlett corrected likelihood ratio statistic. An obvious choice of B is

$$B = \frac{E[W^*(\psi)]}{d}. \tag{16}$$

Bartlett [10] showed that the Bartlett corrected likelihood ratio statistic is also asymptotically χ^2_d distributed and it has rate of convergence $O(n^{-2})$. Therefore, it is an extremely accurate method. Nevertheless, except in a few well-defined problem, $E[W^*(\psi)]$ is very difficult to obtain which hinders the use of this method in applied statistics. A review of the Bartlett corrected likelihood ratio method can be found in Barndorff-Nielsen and Cox [6].

Although, mathematically, the explicit closed form of B, or even an asypmptotic expansion of B, is difficult to obtain, we propose the following algorithmic way to obtain $E[W(\psi)]$ numerically, and hence, an estimated B.

Given: $(x_1, \ldots, x_n)$ is a sample of size n from a distribution with known probability density function $f(\cdot; \theta)$.

Interest: Inference concerning $\psi = \psi(\theta)$.

Have: Overall maximum likelihood estimate $\widehat{\theta}$, the constrained maximum likelihood estimate $\widetilde{\theta}$, and the observed likelihood ratio statistic $w(\psi)$.

Step 1: Simulate M samples of data of size n from $f(\cdot; \widetilde{\theta})$.

Step 2: For each set of simulated data, obtain the simulated observed likelihood ratio statistic. As a result, we have $w_1(\psi), \ldots, w_M(\psi)$.

Step 3: Calculate

$$\overline{w}(\psi) = \frac{\sum_{i=1}^{M} w_i(\psi)}{M}, \tag{17}$$

which is an estimate of the mean of the likelihood ratio statistic. Hence, we have $\widehat{B} = \overline{w}(\psi)/d$.

Step 4: The observed simulated Bartlett corrected likelihood ratio statistic is

$$w^*(\psi) = \frac{w(\psi)}{\widehat{B}}, \tag{18}$$

is asymptotically distributed as χ^2_d with fourth order rate of convergence. Thus, the significance function is $p(\psi) = P(\chi^2_d \le w^*(\psi))$, and the $(1-\alpha)100\%$ confidence region of ψ is

$$\left\{\psi : w^*(\psi) \le \chi^2_{d,1-\alpha}\right\}. \tag{19}$$

As a final note on the proposed algorithm, theoretically, the choice of M should be as large as possible. However, the larger M is, the more calculations are required to obtain $\overline{w}(\psi)$. Moreover, the more nuisance parameters exist in the model, the larger M has to be. We recommend to use trial-by-error of M until $\overline{w}(\psi)$ is stablized.

3.2. Applying Likelihood Based Method to Obtain Inference on the Ratio of Two Coefficients of Variation of Two Independent Log Normal Distribution. Let $\overline{Y_i} = \sum_{j=1}^{n_i} Y_{ij}/n_i$ and $(n_i-1)S_i^2 = \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_i})^2$. Then $\overline{Y_i}$ is normally distributed with mean μ_i and variance σ_i^2/n, and $(n_i-1)S_i^2/\sigma_i^2$ is χ^2_{n-1}. Moreover, $\overline{Y_i}$ and $(n_i-1)S_i^2/\sigma_i^2$ are independent. Hence, inference concerning σ_i^2 will be based on $\chi^2_{n_i-1}$. Since τ_i is a function of σ_i^2 only, inference concerning τ_i will be based on $\chi^2_{n_i-1}$. Let $\theta = (\sigma_1^2, \sigma_2^2)$. Then the likelihood function for θ can be written as

$$\begin{aligned}\ell(\theta) &= \ell\left(\sigma_1^2, \sigma_2^2; s_1^2, s_2^2\right) \\ &= -\frac{n_1-1}{2}\log\sigma_1^2 - \frac{(n_1-1)s_1^2}{2\sigma_1^2} - \frac{n_2-1}{2}\log\sigma_2^2 \\ &\quad - \frac{(n_2-1)s_2^2}{2\sigma_2^2}.\end{aligned} \tag{20}$$

It is easy to show that the overall MLE

$$\hat{\theta} = \left(\hat{\sigma}_1^2, \hat{\sigma}_2^2\right) = \left(s_1^2, s_2^2\right). \tag{21}$$

Since our parameter of interest is $\psi = \psi(\theta) = \tau_1/\tau_2$, where $\tau_i = \sqrt{\exp\{\sigma_i^2\} - 1}$, we have

$$\sigma_1^2 = \log\left(\psi^2 \exp\left\{\sigma_2^2\right\} - \psi^2 + 1\right). \tag{22}$$

For a given ψ value, the log-likelihood function in (20) can be expressed as a function of σ_2^2 only, and is

$$\begin{aligned}\ell\left(\sigma_2^2\right) &= \ell\left(\sigma_2^2; s_1^2, s_2^2\right) \\ &\quad - \frac{n_1-1}{2}\log\left[\log\left(\psi^2\exp\left\{\sigma_2^2\right\} - \psi^2 + 1\right)\right] \\ &\quad - \frac{(n_1-1)s_1^2}{2\left[\log\left(\psi^2\exp\left\{\sigma_2^2\right\} - \psi^2 + 1\right)\right]} \\ &\quad - \frac{n_2-1}{2}\log\sigma_2^2 - \frac{(n_2-1)s_2^2}{2\sigma_2^2}.\end{aligned} \tag{23}$$

Hence, to solve for the constrained MLE $\tilde{\theta} = (\tilde{\sigma}_1^2, \tilde{\sigma}_2^2)$, we have to find $\tilde{\sigma}_2^2$ that maximized (23), and then $\tilde{\sigma}_1^2 = \log(\psi^2\exp\{\tilde{\sigma}_2^2\} - \psi^2 + 1)$. Once we have both the overall and constrained MLEs, we can obtain the observed likelihood ratio statistic $w(\psi)$ as given in (13). Therefore, the significance function is $p(\psi) = P(\chi_1^2 \leq w(\psi))$. Moreover, by applying the algorithm given in the previous section, we can also obtain the observed simulated modified likelihood ratio statistic $w^*(\psi)$ and the corresponding significance function is $p(\psi) = P(\chi_1^2 \leq w^*(\psi))$.

4. Real Data Example

To illustrate the application of the methods discussed in this paper, we revisit the example discussed in Nam and Kwon [4]. Faupel-Badger et al. [11] compare concentrations of estrogen metabolites by RIA with the concentrations obtained using a novel and high-performance liquid chromatography-tandem mass spectrometry (LC-MS/MS). The 10% blinded quality control samples were used for assessment of quality control of the laboratory assay. Partial summary of data were presented in Nam and Kwon [4] and we have

Table 1: 95% confidence interval for ψ for the experiment by Faupel-Badger et al. [11].

Method	95% confidence interval for ψ
Wald type	(2.1798, 3.8620)
Fieller type	(2.2705, 3.9992)
Log method	(2.2867, 3.9907)
Likelihood ratio	(2.2776, 4.0126)
Bartlett correction	(2.2770, 4.0138)

$$\begin{aligned} n_1 &= 48, \\ \hat{\sigma}_1^2 &= (0.1687)^2, \\ n_2 &= 53, \\ \hat{\sigma}_2^2 &= (0.0562)^2 \end{aligned} \tag{24}$$

where the first sample is taken from RIA, and the second sample is taken from LC-MS/MS. Table 1 records the 95% confidence interval for the ratio of the two coefficients of variation assuming that the data are obtained from independent lognormal distributions obtained by the methods discussed in this paper. Note that the MOVER method is identical to the Log method and Hasan and Krishnamoorthy [5] showed that results from the improved version of the MOVER method are still similar to those obtained by the Log method. Hence, both the MOVER method and its improved version are not included in the calculations. Except for the Wald type, the intervals obtained in Nam and Kwon [4] seem to be close to each other. Notice that the results from the Fieller type are different from that reported in Nam and Kwon [4]. Moreover, we observed that the likelihood ratio method and the proposed Bartlett correction method seem to be different from the other methods by having a larger upper confidence limit.

With the above observation, it is of interest to compare the accuracy of the methods discussed in this paper, especially when the sample size is small.

5. Simulation Studies

To compare the accuracy of the methods discussed in this paper, simulations studies are performed. The parameters settings are given in Table 2. Other settings have also been calculated but not reported because the results are very similar to those presented. However, they are available upon request. Since we are interested in developing a method that is accurate even for small sample sizes, hence the chosen sample sizes in the simulations studies are relatively small.

Table 2: Parameters settings for simulation studies.

Study	μ_1	σ_1	μ_2	σ_2	n_1	n_2
1	0.6	0.1	3.0	0.5	5	5
2					10	20
3					15	25
4					20	10
5	1.1	0.2	0.8	0.4	5	5
6					10	20
7					15	25
8					20	10
9	2.5	1.2	3.0	0.7	5	5
10					10	20
11					15	25
12					20	10
13	5.0	0.7	6.0	1.4	5	5
14					10	20
15					15	25
16					20	10

For each study, we obtain 10,000 simulated samples. Theoretically, M should be as large as possible because we want to use $\overline{w}(\psi)$ to be the estimate of $E[W(\psi)]$. However, numerically, we have N simulated samples, and for each simulated sample, we have to do M simulations to obtain $\overline{w}(\psi)$. For these simulation studies, we use $M = 500$. For each simulated sample, we compute the 95% confidence interval obtained by the methods discussed in this paper. Table 3 reported the percentage of samples where the true ψ is less than the lower 95% confidence limit (le), is within the 95% confidence interval (cc), and is greater than the upper 95% confidence limit (ue). The nominal values are 2.5%, 95% and 2.5%, respectively.

From Table 3, the three methods discussed in Nam and Kwon [4] do not give satisfactory coverage, especially when the sample sizes are small. The coverage of the likelihood ratio method is improving when the sample sizes increase and, in general, it has asymmetric errors. Nevertheless the proposed simulated Bartlett corrected likelihood ratio method is extremely accurate even when the sample sizes are as small as 5.

Table 3

(a) Empirical coverage rate for the simulation studies 1 to 8

Study	Method	le	cc	ue
1	Wald type	0.08	87.78	12.14
	Fieller type	3.48	95.50	1.02
	Log method	6.28	88.51	5.21
	Likelihood ratio	3.12	93.39	3.49
	Bartlett correction	2.40	94.90	2.70
2	Wald type	0.20	89.31	10.49
	Fieller type	0.82	94.01	5.17
	Log method	1.94	92.71	5.35
	Likelihood ratio	2.22	94.78	3.00
	Bartlett correction	1.92	95.29	2.79
3	Wald type	0.68	92.92	6.40
	Fieller type	3.48	95.50	1.02
	Log method	3.93	93.25	2.82
	Likelihood ratio	2.77	94.57	2.66
	Bartlett correction	2.53	94.85	2.27
4	Wald type	1.44	94.14	4.42
	Fieller type	7.81	91.99	0.20
	Log method	7.04	91.46	1.50
	Likelihood ratio	3.15	94.23	2.62
	Bartlett correction	2.88	94.85	2.27
5	Wald type	0.05	87.47	12.48
	Fieller type	2.99	95.28	1.73
	Log method	5.88	88.45	5.67
	Likelihood ratio	3.13	93.39	3.48
	Bartlett correction	2.42	94.90	2.68
6	Wald type	0.13	89.18	10.69
	Fieller type	0.67	93.59	5.74
	Log method	1.84	92.58	5.58
	Likelihood ratio	2.25	94.75	3.01
	Bartlett correction	2.00	95.23	2.77
7	Wald type	0.57	92.73	6.70
	Fieller type	2.87	95.26	1.87
	Log method	3.68	93.20	3.12
	Likelihood ratio	2.70	94.66	2.64
	Bartlett correction	2.52	94.90	2.58
8	Wald type	1.34	94.14	4.52
	Fieller type	7.36	92.28	0.36
	Log method	6.68	91.63	1.69
	Likelihood ratio	3.15	94.34	2.49
	Bartlett correction	2.88	94.83	2.29

(b) Empirical coverage rate for the simulation studies 9 to 16

Study	Method	le	cc	ue
9	Wald type	0.00	82.10	17.90
	Fieller type	0.01	98.88	1.11
	Log method	1.27	90.26	8.47
	Likelihood ratio	2.99	93.61	3.40
	Bartlett correction	2.40	94.94	2.66

6. Testing Homogeneity of Coefficients of Variation from k Independent Lognormal Distributions

For k samples from independent lognormal(μ_i, σ_i) distribution, the required log-likelihood function can be written as

$$\begin{aligned}\ell(\theta) &= \ell\left(\sigma_1^2, \ldots, \sigma_k^2; s_1^2, \ldots, s_k^2\right) \\ &= \sum_{i=1}^{k}\left[-\frac{n_i - 1}{2}\log\sigma_i^2 - \frac{(n_i - 1)\, s_i^2}{2\sigma_i^2}\right]\end{aligned} \tag{25}$$

(b) Continued.

Study	Method	le	cc	ue
10	Wald type	0.00	83.91	16.09
	Fieller type	0.00	90.18	9.82
	Log method	0.07	90.86	9.07
	Likelihood ratio	2.04	94.82	3.14
	Bartlett correction	1.92	95.24	2.84
11	Wald type	0.00	89.36	10.64
	Fieller type	0.00	96.70	3.30
	Log method	0.96	93.93	5.11
	Likelihood ratio	2.75	94.59	2.66
	Bartlett correction	2.48	94.97	2.55
12	Wald type	0.02	92.69	7.29
	Fieller type	0.47	99.16	0.37
	Log method	3.62	93.86	2.52
	Likelihood ratio	3.17	94.34	2.49
	Bartlett correction	2.96	94.71	2.33
13	Wald type	0.00	79.33	20.67
	Fieller type	2.01	97.99	0.00
	Log method	10.56	88.87	0.57
	Likelihood ratio	3.52	94.54	2.94
	Bartlett correction	2.68	95.03	2.29
14	Wald type	0.01	91.26	8.73
	Fieller type	0.64	99.32	0.04
	Log method	3.15	94.73	2.12
	Likelihood ratio	2.44	94.69	2.87
	Bartlett correction	2.10	95.19	2.71
15	Wald type	0.59	93.88	5.53
	Fieller type	5.44	94.56	0.00
	Log method	6.29	93.43	0.28
	Likelihood ratio	3.01	94.48	2.51
	Bartlett correction	2.84	94.83	2.33
16	Wald type	1.60	94.18	4.22
	Fieller type	12.16	87.84	0.00
	Log method	10.64	89.36	0.00
	Likelihood ratio	3.01	94.48	2.51
	Bartlett correction	2.95	94.71	2.34

where s_i^2 is the unbiased sample variance estimate of the i^{th} sample given in Section 3. It is well-known that the overall MLE is

$$\hat{\theta} = \left(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_k^2\right) = \left(s_1^2, \ldots, s_k^2\right). \tag{26}$$

The aim is to test

$$\begin{aligned} H_0&: \tau_1 = \cdots = \tau_k = \tau \\ vs\ H_a&: \text{not all coefficients of variation are the same,} \end{aligned} \tag{27}$$

which, in this case, is the same as testing

$$\begin{aligned} H_0&: \sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2 \\ vs\ H_a&: \text{not all variances are the same.} \end{aligned} \tag{28}$$

Therefore, when H_0 is true, the log-likelihood function can be re-written in terms of σ^2 and is

$$\begin{aligned} \ell\left(\sigma^2\right) &= \ell\left(\sigma^2; s_1^2, \ldots, x_k^2\right) \\ &= \sum_{i=1}^{k}\left[-\frac{n_i - 1}{2}\log\sigma^2 - \frac{\left(n_i - 1\right)s_i^2}{2\sigma^2}\right], \end{aligned} \tag{29}$$

and the constrained MLE is

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{k}\left(n_i - 1\right)s_i^2}{\sum_{i=1}^{k}\left(n_i - 1\right)}, \tag{30}$$

which is the usual pooled variance estimate. The observed likelihood ratio statistic is

$$w = 2\left[\ell\left(\hat{\theta}\right) - \ell\left(\tilde{\sigma}^2\right)\right], \tag{31}$$

which is asymptotically distributed as χ_{k-1}^2. Hence, the observed simulated Bartlett corrected likelihood ratio statistic is

$$w^* = \frac{w}{\overline{w}/\left(k - 1\right)}, \tag{32}$$

where $\overline{w}$ is obtained by the algorithm given in Section 2.

Simulation studies are performed to compare the accuracy of the likelihood ratio method and the simulated Bartlett corrected likelihood ratio method. In particular, three samples of data from lognormal(μ_i, σ) distribution are generated. w is calculated and w^* is also the calculation with $M = 1000$. We repeat this process $N = 10{,}000$. The proportion of samples that have p-values less than 5% is reported in Table 4 for various sample sizes. The choice of μ_i is not important because it does not involve in any of the calculations and, hence, we take it to be 0. Different choices of σ result in similar results and are not reported, but they are available upon request. Table 4 reported the cases $\mu_i = 0$ and $\sigma = 1$. When sample sizes are small, the likelihood ratio method does not give satisfactory results, but it is improving when the sample sizes increase. The simulated Bartlett corrected likelihood ratio method consistently gives extremely accurate result even when the sample sizes are small.

7. Conclusion

The lognormal distribution has been frequently used for modeling positive valued right skewed data, which commonly arise in health and medical research. In this paper, we proposed a simulated Bartlett corrected likelihood ratio approach to obtain inference concerning the ratio of two coefficients of variation for lognormal distribution. Simulation studies show that the proposed Bartlett correction method is extremely accurate even when the sample size is small. Moreover, the proposed proposed Bartlett correction method is extended to test homogeneity of k coefficients of variation from independent lognormal distributions.

Table 4: The proportion of samples rejected at 5% level of significance for testing $H_0 : \tau_1 = \tau_2 = \tau_3 = \tau$ when data are generated from lognormal(0, 1) distribution.

	Proportion of samples rejected at $\alpha = 0.05$	
(n_1, n_2, n_3)	Likelihood ratio	Bartlett correction
(5, 5, 5)	0.0658	0.0512
(5, 10, 15)	0.0646	0.0524
(10, 10, 10)	0.0565	0.0495
(10, 15, 20)	0.0529	0.0487
(50, 50, 50)	0.0544	0.0523

Data Availability

The data set for compared concentrations of estrogen metabolites by RIA with the concentrations obtained using a novel and high-performance liquid chromatography-tandem mass spectrometry (LC-MS/MS) is from previously reported in Faupel-Badger et al. [11], which has been cited. This data set was further analyzed in Nam and Kwon [4], which was also cited in the manuscript. The other numerical examples in the submitted paper are based on simulation studies, which is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] J. Aitchison and J. A. C. Brown, *The Lognormal Distribution*, Cambridge University Press, Cambridge, UK, 1957.

[2] E. L. Crow and K. Simizu, *Lognormal Distributions. Theory and Application*, Marcel Dekker, New York, NY, USA, 1988.

[3] T. R. Fears, R. G. Ziegler, J. L. Donaldson et al., "Reproducibility studies and interlaboratory concordance for androgen assays of male plasma hormone levels," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 11, no. 8, pp. 785–789, 2002.

[4] J. Nam and D. Kwon, "Inference on the ratio of two coefficients of variation of two lognormal distributions," *Communications in Statistics—Theory and Methods*, vol. 46, no. 17, pp. 8575–8587, 2017.

[5] M. S. Hasan and K. Krishnamoorthy, "Improved confidence intervals for the ratio of coefficients of variation of two lognormal distributions," *Journal of Statistical Theory and Applications*, vol. 16, no. 3, pp. 345–353, 2017.

[6] O. E. Barndorff-Nielsen and D. R. Cox, *Inference and Asymptotics*, Chapman and Hall, New York, NY, USA, 1994.

[7] D. A. S. Fraser, "P-values: The insight to modern statistical inference," *Annual Review of Statistics and Its Application*, vol. 4, pp. 1–14, 2017.

[8] O. E. Barndorff-Nielsen, "Inference on full or partial parameters based on the standardized signed log likelihood ratio," *Biometrika*, vol. 73, no. 2, pp. 307–322, 1986.

[9] O. E. Barndorff-Nielsen, "Modified signed log likelihood ratio," *Biometrika*, vol. 78, no. 3, pp. 557–563, 1991.

[10] M. S. Bartlett, "Properties of sufficiency and statistical tests," *Proceedings of the Royal Society A Mathematical, Physical and Engineering Sciences*, vol. 160, no. 901, pp. 268–282, 1937.

[11] J. M. Faupel-Badger, B. J. Fuhrman, X. Xu et al., "Comparison of liquid chromatography-tandem mass spectrometry, RIA, and ELISA methods for measurement of urinary estrogens," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 19, no. 1, pp. 292–300, 2010.

Hindawi
Journal of Probability and Statistics
Volume 2019, Article ID 9750538, 5 pages
https://doi.org/10.1155/2019/9750538

Review Article

A Comparison of Mean-Based and Quantile Regression Methods for Analyzing Self-Report Dietary Intake Data

Michelle L. Vidoni,[1] Belinda M. Reininger,[2] and MinJae Lee[1,3]

[1] *Biostatistics, Epidemiology, and Research Design (BERD) Core, Center for Clinical and Translational Sciences (CCTS), The University of Texas Health Science Center at Houston, 6410 Fannin, Houston, TX 77030, USA*
[2] *Health Promotion & Behavioral Sciences, Hispanic Health Research Center, The University of Texas School of Public Health Brownsville Regional Campus, One West University Blvd., Brownsville, TX 78520, USA*
[3] *Division of Clinical and Translational Sciences, Department of Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, 6410 Fannin, Houston, TX 77030, USA*

Correspondence should be addressed to MinJae Lee; minjae.lee@uth.tmc.edu

Received 17 July 2018; Accepted 12 February 2019; Published 3 March 2019

Guest Editor: Min Zhang

In mean-based approaches to dietary data analysis, it is possible for potentially important associations at the tails of the intake distribution, where inadequacy or excess is greatest, to be obscured due to unobserved heterogeneity. Participants in the upper or lower tails of dietary intake data will potentially have the greatest change in their behavior when presented with a health behavior intervention; thus, alternative statistical methods to modeling these relationships are needed to fully describe the impact of the intervention. Using data from *Tu Salud ¡Si Cuenta! (Your Health Matters!) at Home Intervention*, we aimed to compare traditional mean-based regression to quantile regression for describing the impact of a health behavior intervention on healthy and unhealthy eating indices. The mean-based regression model identified no differences in dietary intake between intervention and standard care groups. In contrast, the quantile regression indicated a nonconstant relationship between the unhealthy eating index and study groups at the upper tail of the unhealthy eating index distribution. The traditional mean-based linear regression was unable to fully describe the intervention effect on healthy and unhealthy eating, resulting in a limited understanding of the association.

1. Introduction

Many health behavior interventions focus on positive lifestyle changes in the areas of increasing physical activity and healthy diets. Adopting these behavior changes can prevent or reduce the negative health consequences of obesity in minority US populations. Mexican Americans are particularly prone to physical inactivity and poor diets because of lack of fruit and vegetable consumption compared to Non-Hispanic Whites [1, 2]. Despite research showing poorer dietary intake than other ethnic groups, within the Mexican American population there is heterogeneity in healthy and unhealthy food intake [3].

Dietary intake data is typically measured using self-report tools and individual food intake is aggregated into compositional data or patterns to describe overall diets. When the dietary data are analyzed using mean-based approaches, such as ordinary least squares (OLS) regression, potentially important relationships with disease risk at the lower and upper levels of the distribution could be obscured due to unobserved heterogeneity. Participants in the upper or lower tails of dietary intake data, where inadequacy or excess is greatest, will theoretically have the greatest change in their behavior when presented with a health behavior intervention; thus, alternative statistical methods to modeling these relationships are needed to fully describe the impact of the intervention. This is particularly notable in certain populations, such as Mexican Americans, where variation in factors such as acculturation and language influence food choices and adherence to traditional and western diet patterns [3–6].

As an alternative to mean-based regression, quantile regression (QR) was developed by Koenker and Bassett and has primarily been used in the fields of risk management and business [7]. Quantile regression has been extended for handling longitudinal data based on different approaches that account for serial correlations within a subject and has been used as an important alternative to mean-based regression approaches because of its flexibility for modeling nonnormal data, or heterogeneous conditional distributions [8]. QR can model the conditional distribution of the response, not only on the conditional mean, giving the research critical insights when valuable information lies in the tails. Despite QR being computationally intensive and not equipped to handle small data sets, it is more robust to outliers than mean-based regression, where estimates of the conditional mean can be strongly influenced by outliers.

Application of QR to health and behavioral sciences is increasing and could be a valuable statistical tool for health researchers. QR has been used to evaluate the effects of physical activity or dietary intake on varying quantile levels of certain variables, such as BMI [9–12], waist circumference [13], socioeconomic status [14], and risk factors of disease outcomes including health-related scores and biomarker data [15–19]. A limited number of studies have introduced a QR-based approach specifically applied to behavioral data [20–22]. Yet, there is limited research focusing on how to use and apply QR results to improve behavioral interventions and maintenance of behavior change over time by possibly addressing the upper and lower tails of the population distribution differently.

The goal of this review was to compare traditional mean-based linear regression with QR through the illustration of their applications to real data from the behavioral intervention study aimed at improving healthy eating and to demonstrate the usefulness of QR in fully describing the relationships.

2. Linear Quantile Mixed Effect Regression

Let y_{it} be the measurement for the i-th subject ($i = 1, \ldots, n$) at time t ($t = 1, \ldots, n_i$), then we define a linear mixed effect regression model as

$$\mathrm{y}_{it} = \boldsymbol{x}_{it}^{T}\boldsymbol{\beta} + \boldsymbol{z}_{it}^{T}\boldsymbol{\gamma}_i + \varepsilon_{it}, \qquad (1)$$

where $\boldsymbol{x}_{it}$ is a vector of p covariates at t, $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of regression parameters, and the correlation among the observations within the i-th subject is induced by the subject-level residuals, i.e., $q \times 1$ vector $\boldsymbol{\gamma}_i$ and an associated vector $\boldsymbol{z}_{it}$ for q random effect variables. The error term can be defined as $\boldsymbol{z}_{it}^{T}\boldsymbol{\gamma}_i + \varepsilon_{it}$, where random errors for individual records, ε_{it}, are independent of each other. We assume that linear quantile mixed models are determined based on the asymmetric Laplace distribution (ALD) [23], which has a good performance on data generated from many error distributions, and a relationship with the L_1-norm objective function [7]. Let a response variable y be an ALD, denoted $\mathrm{ALD}(\mu, \sigma, \tau)$, then we can define a probability density function,

$$f(y \mid \mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\}, \qquad (2)$$

where $0 < \tau < 1$ is the skewness parameter, μ is the location parameter, σ is the scale parameter, and a loss function $\rho_\tau(v) = (\tau - I(v \le 0))$ represents the contribution by residuals v. Assuming the location parameter is $\mu_{\tau,it} = \boldsymbol{x}_{it}^{T}\boldsymbol{\beta}_\tau + \boldsymbol{z}_{it}^{T}\boldsymbol{\gamma}_i$, a quantile regression model related to the τ-th quantile of a response variable y_{it}, conditional on $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$, has the form:

$$q_\tau(y_{it}) = \boldsymbol{x}_{it}^{T}\boldsymbol{\beta}_\tau + \boldsymbol{z}_{it}^{T}\boldsymbol{\gamma}_i + \varepsilon_{\tau,it}, \quad 0<\tau<1, \qquad (3)$$

where $\boldsymbol{\beta}_\tau$ is a vector of quantile-specific regression parameters corresponding to the coefficient $\boldsymbol{\beta}$ in a linear regression model (1) and $\varepsilon_{\tau,it} \sim \mathrm{ALD}(0, \sigma, \tau)$, which are also dependent on τ. The objective function for y_{it} for fixed τ is expressed as

$$Q_n(\boldsymbol{\beta}_\tau) = \sum_{i=1}^{n}\sum_{t=1}^{n_i}\rho_\tau\left(y_{it} - \boldsymbol{x}_{it}^{T}\boldsymbol{\beta} - \boldsymbol{z}_{it}^{T}\boldsymbol{\gamma}_i\right). \qquad (4)$$

We can estimate quantile-specific regression parameters that minimize the objective function above. As we assume $y_{it} \sim \mathrm{ALD}(\mu_{it}, \sigma, \tau)$, ALD is determined as a scale mixture of normal distribution based on Laplace distribution with the skewness parameter τ that is treated here as a quantile level. Then a likelihood for y_{it} at τ-th quantile can be expressed as

$$L(\boldsymbol{\beta}, \sigma \mid y_{it}, \tau) = \frac{\tau(1-\tau)}{\sigma^N} \exp\left\{-\sum_{i=1}^{n}\sum_{t=1}^{n_i}\rho_\tau\left(\frac{y_{it} - \boldsymbol{x}_{it}^{T}\boldsymbol{\beta} - \boldsymbol{z}_{it}^{T}\boldsymbol{\gamma}_i}{\sigma}\right)\right\}. \qquad (5)$$

If σ is considered a nuisance parameter, then the maximization of this likelihood above is equivalent to the minimization of the objective function of quantile regression (4) defined above. More details regarding estimation process are available elsewhere [8].

3. Example

3.1. Tu Salud ¡Si Cuenta! (Your Health Matters!) at Home Intervention. The behavioral data used in the current study were from the *Tu Salud ¡Si Cuenta! (Your Health Matters!) at Home Intervention*. One of the main objectives of this randomized control trial was to increase participant intake of healthy foods and decrease unhealthy food intake through exposure to community health workers delivering a behavioral modification intervention. The study was conducted in the Texas Rio Grande Valley area and included participants who were Mexican American adults, aged 18-75 years, and enrolled in the ongoing Cameron County Hispanic Cohort [1, 24] . Participants were randomly selected and randomized into either the intervention or standard care group from June 2010 to April 2013. The intervention group received up to six monthly community health work home visits in the first 6 months of the intervention, which included lifestyle change education, motivation, and support. No other intervention elements, other than that equivalent to the standard care group, were offered during the last 6 months of the trial. The standard care group participants were potentially exposed to a community-wide physical activity and healthy diet campaign across the 12 months. Data were collected at baseline, 6- and 12-month follow-ups.

Participants completed a dietary intake questionnaire that asked if yesterday they had eaten 20 commonly and culturally appropriate foods and how many times with the following responses available: no, once, twice, three times, four times, and five or more times [25, 26]. Responses were summed into Healthy and Unhealthy Eating Indices (HEI and UNHEI, respectively). The HEI score was comprised responses to the 10 healthy food items (baked or grilled fish, turkey or chicken; eggs; beans; fruit; fruit juice; orange vegetables; other vegetables; salad; whole grain breads; and whole grain cereals) with a possible response range from 0 to 50. The UNHEI was composed of the responses to the 9 unhealthy food items (baked goods; french fries or chips; fried meat; frozen desserts; red and processed meats; nonchocolate candy; regular sodas; sweetened or sports drinks; and white bread) with a possible range from 0 to 45 [27]. Both HEI and UNHEI scores appeared to be well-approximated by a normal distribution.

3.2. Quantile Regression and Mean-Based Regression. To assess intervention effect on healthy and unhealthy eating, a multivariable longitudinal QR and mean-based model were conducted based on the linear mixed effect model equation below.

$$\begin{aligned} \mathrm{y}_{it} &= \alpha + \beta_1 \mathrm{x}_{1i} \\ &\quad + \beta_{21}\mathrm{v}_{1it} + \beta_{22}\mathrm{v}_{2it} + \beta_{31}\mathrm{x}_{1i}\mathrm{v}_{1it} + \beta_{32}\mathrm{x}_{1i}\mathrm{v}_{2it} + \boldsymbol{u}_{it}^T\boldsymbol{\delta} + \gamma_i + \varepsilon_{it}, \end{aligned} \tag{6}$$

where the index score, y_{it}, can be either the HEI or UNHEI measurement for the i-th participant ($i = 1, \dots, n$) at visit t ($t = 1, 2, 3$) and a binary variable x_1 for study group (x_1=1 if intervention) and v_{1i} and v_{2i} are dummy variables for two follow-up visits, i.e., month 6 (visit 2) and month 12 (visit 3), respectively. Interaction terms between study group and follow-up visits were included in the model to obtain estimates of the intervention effect at each time point. $\boldsymbol{u}_{it}$ is a vector of a set of potential confounders that were adjusted for in the model (i.e., gender, age, diabetes, marital status, years in school, employment status, type of insurance, generation, and preferred language) and $\boldsymbol{\delta}$ is an associated parameter vector. We also considered a random intercept by including an error term γ_i for the i-th subject. We used *lqmm* R package [8] for QR models and SAS *proc mixed* for mean-based models.

3.3. Results. There were 500 participants randomized to either the standard care or intervention groups, n=250 respectively. At baseline, the mean HEI score was 6.6 (standard deviation (SD)=3.3) for the standard care group and 6.9 (SD=3.5) for the intervention group. The mean UNHEI score for the standard care group was 5.4 (SD=3.4) and for the intervention group was 5.6 (SD=3.6).

Results from QR and mean-based regression are presented in Figure 1. The red line indicates estimated beta coefficients based on mean-based model for the effect of the study group at each time point, showing slight differences (i.e., beta coefficient <0.4) in mean HEI and mean UNHEI between intervention and standard care groups at baseline and follow-ups.

With regard to HEI, the results for QR and mean-based regression do not substantially differ. In contrast, the QR results indicate a nonconstant relationship between unhealthy eating and study groups at the upper tail of the distribution of the UNHEI. At baseline, the association between the distribution of UNHEI scores and study groups is not constant, as the intervention group is more likely to be in the upper tail of the UNHEI distribution at the start of the study. At month 6, the effect of the intervention is inconsistent across the UNHEI distribution. For example, at the upper tail of the UNHEI distribution the intervention group had higher UNHEI scores, yet around the quantile level τ=0.05 and 0.75 the intervention group reported lower UNHEI scores than the control group. The strength of the association in the upper tail of the distribution is attenuated at 6 months compared to baseline. More strikingly, at the 12-month follow-up QR suggests that there is an increase in unhealthy food intake in intervention group compared to control for the participants in the upper tail of the UNHEI data distribution.

4. Discussion

Mean-based regression results showed minimal differences in the healthy eating index at any visit between intervention and standard care groups, likewise for the unhealthy eating index. These results would lead a researcher to incorrectly assume that the intervention failed to increase intake of healthy foods or decrease unhealthy food intake or possibly conclude that the reasons for the lack of change might not be due to the intervention itself but to information bias or environmental changes in the community based intervention.

In contrast, the results of the QR highlight a different relationship between the study groups and outcomes. The estimated coefficients were not constant across the distribution of the UNHEI outcome at baseline and follow-ups. These results may indicate a baseline imbalance in the UNHEI outcome, which under mean-based regression would have not been identified, and approaches to adjust for the imbalance should be considered. Likewise at the 6-month follow-up, the protective effect of the intervention would have also been ignored using mean-based methods. The QR results for the unhealthy index at the 12-month follow-up identified an inconsistent relationship between study group and UNHEI. At the lower tail of UNHEI, the intervention was protective, then this relationship reversed at the upper tail of UNHEI. Overall, there was little difference in the UNHEI between intervention and standard care groups, except at the upper tail of the UNHEI distribution. This indicates purely mean-based approach may not be appropriate for evaluating the effect of the intervention on dietary uptake behaviors in populations with unobserved heterogeneity.

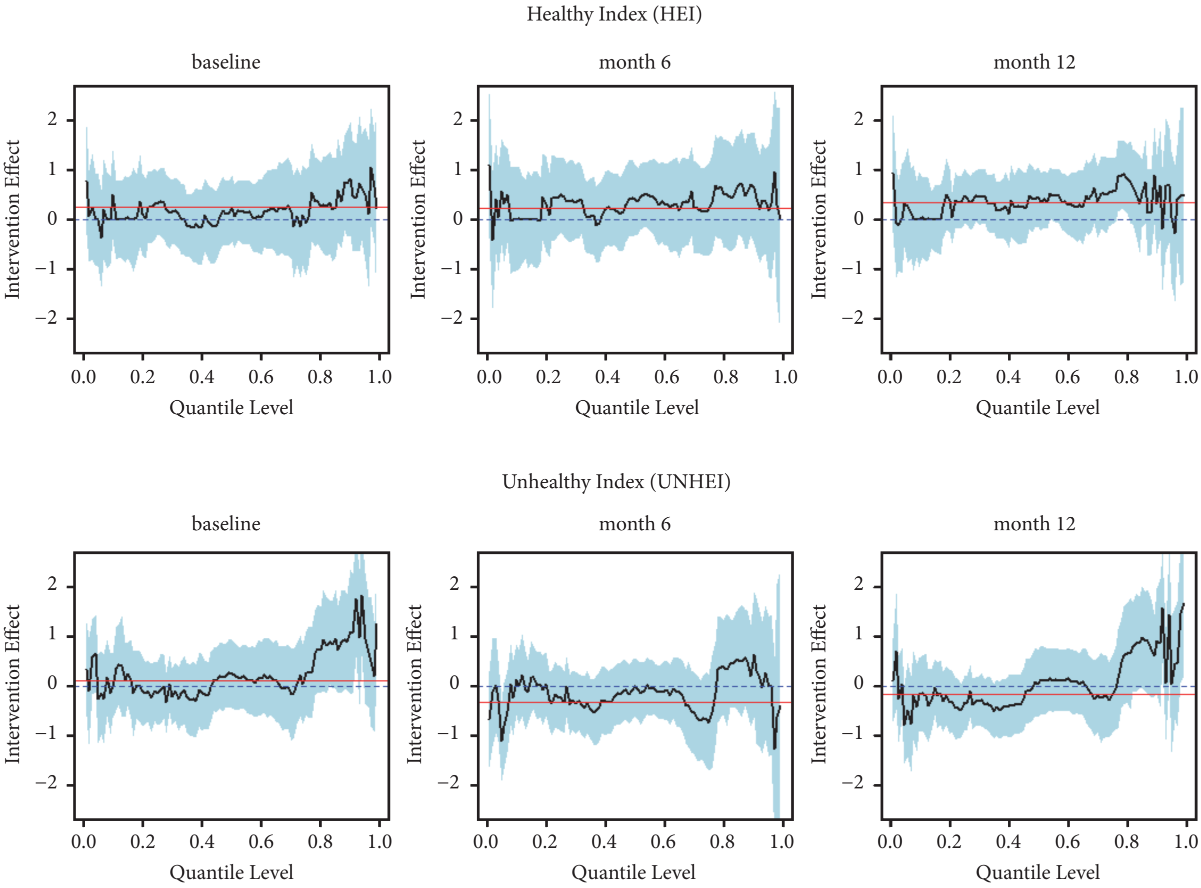


Figure 1: Estimated adjusted∗ parameter of x_1 for study group (intervention vs. standard care) at each visit based on mean-based regression (red line) and quantile regression (black line with 95% confidence limits) by quantile levels of healthy (HEI) and unhealthy index (UNHEI). ∗Adjusted for gender, age, diabetes, marital status, years in school, employment status, type of insurance, generation, and preferred language.

5. Conclusions

The traditional mean-based linear regression was unable to fully describe the relationship between healthy and unhealthy eating and the intervention, resulting in a limited understanding of the intervention effect. Use of quantile regression identified a different relationship by modeling the coefficients across the distribution of the outcome resulting in a more complete picture of the association. These findings from the quantile regression results could be applied towards developing more effective behavioral intervention trials in heterogeneous populations.

Disclosure

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH CTSA or NIMHD or UTCO.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to acknowledge and thank Dr. Belinda Reininger for her guidance and support on demonstrating the statistical approaches through the behavioral data observed from the intervention program, the *Tu Salud ¡Si Cuenta! (Your Health Matters!) at Home Intervention,* which was supported by the UT Health Clinical and Translational Science Award (UL1 TR000371), NIH/National Institute on Minority Health and Health Disparities (MD000170 P20), and the Texas Department of State Health Services funding for University of Texas Community Outreach (UTCO). The authors would like to recognize the support provided by the Biostatistics/Epidemiology/Research Design (BERD) component of the Center for Clinical and Translational Sciences (CCTS) at the UT Health Science Center at Houston, which is mainly funded by the NIH Centers for Translational Science Award (NIH CTSA), grant UL1 RR024148.

References

[1] B. M. Reininger, L. Mitchell-Bennett, M. Lee et al., “Tu Salud,¡ Si Cuenta!: exposure to a community-wide campaign and

its associations with physical activity and fruit and vegetable consumption among individuals of mexican descent," *Social Science and Medicine*, vol. 143, pp. 98–106, 2015.

[2] B. M. Reininger, J. Wang, S. P. Fisher-Hoch, A. Boutte, K. Vatcheva, and J. B. McCormick, "Non-communicable diseases and preventive health behaviors: a comparison of Hispanics nationally and those living along the US-Mexico border Health behavior, health promotion and society," *BMC Public Health*, vol. 15, no. 1, article 564, 2015.

[3] B. M. Reininger, M. Lee, R. Jennings, A. Evans, and M. Vidoni, "Healthy eating patterns associated with acculturation, sex and BMI among Mexican Americans," *Public Health Nutrition*, vol. 20, no. 7, pp. 1267–1278, 2017.

[4] K. J. Duffey, P. Gordon-Larsen, G. X. Ayala, and B. M. Popkin, "Birthplace is associated with more adverse dietary profiles for US-born than for foreign-born Latino adults," *Journal of Nutrition*, vol. 138, no. 12, pp. 2428–2435, 2008.

[5] I. B. Ahluwalia, E. S. Ford, M. Link, and J. C. Bolen, "Acculturation, weight, and weight-related behaviors among Mexican Americans in the United States," *Ethnicity and Disease*, vol. 17, no. 4, pp. 643–649, 2007.

[6] J. K. Montez and K. Eschbach, "Country of birth and language are uniquely associated with intakes of fat, fiber, and fruits and vegetables among Mexican-American women in the United States," *Journal of the Academy of Nutrition and Dietetics*, vol. 108, no. 3, pp. 473–480, 2008.

[7] R. Koenker and G. Bassett Jr., "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[8] M. Geraci, "Linear quantile mixed models: the lqmm package for laplace quantile regression," *Journal of Statistical Software*, vol. 57, no. 13, pp. 1–29, 2014.

[9] D. A. Amugsi, Z. T. Dimbuene, P. Bakibinga, E. W. Kimani-Murage, T. N. Haregu, and B. Mberu, "Dietary diversity, socioeconomic status and maternal body mass index (BMI): quantile regression analysis of nationally representative data from Ghana, Namibia and Sao Tome and Principe," *BMJ Open*, vol. 6, no. 9, Article ID e012615, 2016.

[10] M. Bottai, E. A. Frongillo, X. Sui et al., "Use of quantile regression to investigate the longitudinal association between physical activity and body mass index," *Obesity*, vol. 22, no. 5, pp. E149–E156, 2014.

[11] S. Azagba and M. F. Sharaf, "Fruit and vegetable consumption and body mass index: a quantile regression approach," *Journal of Primary Care & Community Health*, vol. 3, no. 3, pp. 210–220, 2012.

[12] J. A. Mitchell, R. R. Pate, V. España-Romero, J. R. O'Neill, M. Dowda, and P. R. Nader, "Moderate-to-vigorous physical activity is associated with decreases in body mass index from ages 9 to 15 years," *Obesity*, vol. 21, no. 3, pp. E280–E286, 2013.

[13] J. A. Mitchell, M. Dowda, R. R. Pate et al., "Physical activity and pediatric obesity: a quantile regression analysis," *Medicine and Science in Sports and Exercise*, vol. 49, no. 3, pp. 466–473, 2017.

[14] Ø. Seippel, "Physical exercise and social inequality in Norway – A comparison of OLS and quantile regression analysis," *European Journal for Sport and Society*, vol. 12, no. 4, pp. 355–376, 2016.

[15] A. D'Silva, P. A. Gardiner, T. Boyle, D. G. Bebb, S. T. Johnson, and J. K. Vallance, "Associations of objectively assessed physical activity and sedentary time with health-related quality of life among lung cancer survivors: A quantile regression approach," *Lung Cancer*, vol. 119, pp. 78–84, 2018.

[16] Z. Wang, P. Gordon-Larsen, A. M. Siega-Riz et al., "Sociodemographic disparity in the diet quality transition among Chinese adults from 1991 to 2011," *European Journal of Clinical Nutrition*, vol. 71, no. 4, pp. 486–493, 2017.

[17] L. Liu, "Using multivariate quantile regression analysis to explore cardiovascular risk differences in subjects with chronic kidney disease by race and ethnicity: findings from the US chronic renal insufficiency cohort study," *International Cardiovascular Forum Journal*, 2015.

[18] A. K. Monroe, T. T. Brown, C. Cox, S. M. Reynolds, D. J. Wiley, F. J. Palella et al., "Physical activity and its association with insulin resistance in multicenter AIDS cohort study men," *AIDS Research and Human Retroviruses*, vol. 31, no. 12, pp. 1250–1256, 2015.

[19] E. Verly, J. Steluti, R. M. Fisberg, and D. M. L. Marchioni, "A quantile regression approach can reveal the effect of fruit and vegetable consumption on plasma homocysteine levels," *PLoS ONE*, vol. 9, no. 11, Article ID e111619, 2014.

[20] J. N. Variyam, J. Blaylock, and D. Smallwood, "Characterizing the distribution of macronutrient intake among U.S. Adults: A quantile regression approach," *American Journal of Agricultural Economics*, vol. 84, no. 2, pp. 454–466, 2002.

[21] Y. Wei, Y. Ma, and R. J. Carroll, "Multiple imputation in quantile regression," *Biometrika*, vol. 99, no. 2, pp. 423–438, 2012.

[22] Y. Wei and R. J. Carroll, "Quantile regression with measurement error," *Journal of the American Statistical Association*, vol. 104, no. 487, pp. 1129–1143, 2009.

[23] D. V. Hinkley and N. S. Revankar, "Estimation of the Pareto law from underreported data: a further analysis," *Journal of Econometrics*, vol. 5, no. 1, pp. 1–11, 1977.

[24] S. P. Fisher-Hoch, A. R. Rentfro, J. J. Salinas et al., "Socioeconomic status and prevalence of obesity and diabetes in a Mexican American community, Cameron County, Texas, 2004-2007," *Preventing Chronic Disease*, vol. 7, no. 3, article A53, 2010.

[25] D. M. Hoelscher, R. S. Day, E. S. Lee et al., "Measuring the prevalence of overweight in Texas schoolchildren," *American Journal of Public Health*, vol. 94, no. 6, pp. 1002–1008, 2004.

[26] A. Pérez, D. M. Hoelscher, H. S. Brown, and S. H. Kelder, "Peer reviewed: differences in food consumption and meal patterns in texas school children by grade," *Preventing Chronic Disease*, vol. 4, no. 2, 2007.

[27] C. E. Velazquez, K. E. Pasch, N. Ranjit, G. Mirchandani, and D. M. Hoelscher, "Are adolescents' perceptions of dietary practices associated with their dietary behaviors?" *Journal of the Academy of Nutrition and Dietetics*, vol. 111, no. 11, pp. 1735–1740, 2011.

Hindawi
Journal of Probability and Statistics
Volume 2019, Article ID 8953530, 13 pages
https://doi.org/10.1155/2019/8953530

Research Article

On the Use of Min-Max Combination of Biomarkers to Maximize the Partial Area under the ROC Curve

Hua Ma,[1] Susan Halabi ,[2] and Aiyi Liu[3]

[1] *Merck & Co. Inc., Kenilworth, NJ 07033, USA*
[2] *Department of Biostatistics and Bioinformatics, Box 2717, Duke University Medical Center, Durham, NC 27710, USA*
[3] *Biostatistics and Bioinformatics Branch, Division of Intramural Population Health Research,*
Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, Rockville, MD, USA

Correspondence should be addressed to Susan Halabi; susan.halabi@duke.edu

Received 3 May 2018; Revised 6 December 2018; Accepted 3 January 2019; Published 3 February 2019

Guest Editor: Yichuan Zhao

Background. Evaluation of diagnostic assays and predictive performance of biomarkers based on the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) are vital in diagnostic and targeted medicine. The partial area under the curve (pAUC) is an alternative metric focusing on a range of practical and clinical relevance of the diagnostic assay. In this article, we adopt and extend the min-max method to the estimation of the pAUC when multiple continuous scaled biomarkers are available and compare the performances of our proposed approach with existing approaches via simulations. *Methods*. We conducted extensive simulation studies to investigate the performance of different methods for the combination of biomarkers based on their abilities to produce the largest pAUC estimates. Data were generated from different multivariate distributions with equal and unequal variance-covariance matrices. Different shapes of the ROC curves, false positive fraction ranges, and sample size configurations were considered. We obtained the mean and standard deviation of the pAUC estimates through re-substitution and leave-one-pair-out cross-validation. *Results*. Our results demonstrate that the proposed method provides the largest pAUC estimates under the following three important practical scenarios: (1) multivariate normally distributed data for nondiseased and diseased participants have unequal variance-covariance matrices; or (2) the ROC curves generated from individual biomarker are relative close regardless of the latent normality distributional assumption; or (3) the ROC curves generated from individual biomarker have straight-line shapes. *Conclusions*. The proposed method is robust and investigators are encouraged to use this approach in the estimation of the pAUC for many practical scenarios.

1. Introduction

The area under the entire curve (AUC) is one of the most commonly used summary indices in receiver operating characteristic (ROC) analysis and can be interpreted as the average value of sensitivity for all possible values of specificity [1]. The empirical estimate of the AUC is closely related to the Mann-Whitney U statistic for comparing ratings of nondiseased and diseased participants [1]. Although methods based on the AUC have been well developed and widely implemented [2, 3], one of the major limitations of the AUC is that it summarizes the performance over the entire curve, including regions that may not be clinically relevant (e.g., the regions with low specificity levels). The partial area under the ROC curve (pAUC) can be used as a summary index of diagnostic/prognostic accuracy over a certain range of specificity that is of clinical interest [4, 5]. In many applications, tests with false positive rates outside of a particular domain will be of no practical use and hence are irrelevant for evaluating the accuracy of the test. In particular, for a certain disease with low prevalence, the unnecessary follow-up resulting from high false positive rate will burden the health system. There are several proposed methods for analyzing the pAUC [4, 6–10].

When multiple continuous-scaled biomarkers are available in the evaluation of prognostic accuracy, it may be possible to improve the accuracy by combining several biomarkers. The use of linear combination is popular due to its ease of implementation and interpretation. Finding optimal linear combination to maximize the area under the ROC curve has been extensively studied [11–14]. By extending Fisher's discriminant function, Su and Liu [11] first proposed the best linear combination to maximize AUC based on the multivariate normality assumption. Su and Liu's method relies on the strong distributional assumption, and therefore pAUC may have unsatisfactory performance for many practical scenarios when the distributional assumption is not satisfied. Liu et al. [12] provided an approach to construct the best linear combination that can produce the ROC curve dominating any other ROC curves in some particular specificity ranges. However, this approach depends on the distributional assumption about the mean vectors and the specificity range. Therefore, it may fail to be dominant for a particular range of specificity and sensitivity that may be of clinical interest. In addition, this approach involves the calculation of the eigenvector corresponding to the eigenvalue, and thus the stability of this approach depends on the behavior of eigenvector under small perturbation of the corresponding matrix [15].

Under the assumption of generalized linear model, Jin and Lu [13] proved that the combination coefficients from the estimates of logistic regression yielded ROC curve with the highest sensitivity uniformly over the entire range of specificity. Without distributional assumptions on the data, Pepe and Thompson [16] considered maximizing AUC and pAUC through rank-based estimate, i.e., the Mann-Whitney U statistic [1]. They proposed an algorithm to search for optimal linear combinations with number of biomarkers equal to 2. This approach was computationally formidable when the number of biomarkers is greater than or equal to 3 [17]. Hsu and Hsueh [18] and Yu and Park [19] proposed methods to maximize the partial area under the ROC curve based on the multivariate normality assumption.

Liu et al. [20] developed a nonparametric min-max approach that reduces data into two dimensions to maximize the Mann-Whitney statistic of the AUC. This approach is robust against distributional assumptions due to its nonparametric nature and is computationally efficient since the min-max procedure involves searching for only one single coefficient. Although useful, this approach was developed based on the full range of specificity. In many medical areas, the ROC curve is only clinically relevant and of interest when the assay has high specificities. For example, high specificity of an assay is required for screening any healthy population. Similarly, in using diagnostic assay with multiple genes, only high sensitivity and specificity classifiers have clinical utility (Sparano 2015).

We adapt and extend the min-max method to estimating the pAUC when several markers are considered. This article is organized as follows. In Section 2, we provide a thorough review of existing methods that maximize the AUC and pAUC. In Section 3, we extend the min-max combination method to the optimization of the pAUC and discuss the leave-one-pair-out (LOPO) cross-validation approach for evaluation of the combination methods based on their accuracy for future observations. In Section 4, we then conduct extensive simulations to investigate the performance of the different combination methods based on their abilities to yield the largest pAUC estimates. In Section 5, two real life examples are presented. We then discuss the results in Section 6 and provide guidelines for practical use of the different approaches.

2. Existing Methods

2.1. Definition. Without loss of generality, we consider the partial area under the ROC curve (pAUC) over the range of high specificity values, i.e.,

$$pAUC_{t_0} = \int_0^{t_0} ROC\left(t\right) dt. \tag{1}$$

In this article, t_0 less than or equal to 0.2, i.e., specificity greater than or equal to 0.8, were considered. This is due to the fact that an assay is unlikely to be used if it has a lower specificity rate.

Let X_i, $i = 1, \ldots, n_1$, and Y_j, $j = 1, \ldots, n_2$, be the biomarker levels for nondiseased and diseased participants. The corresponding empirical estimate of pAUC by utilizing the Mann-Whitney U statistic is

$$\widehat{pAUC} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\left(X_i < Y_j \textit{ and } X_i > Q\left(1 - t_0\right)\right) \tag{2}$$

where $Q(1 - t_0)$ is the $(1 - t_0)$ quantile of the empirical distribution of X.

Assume that we have p diagnostic tests or biomarkers on each subject, n_1 nondiseased participants with ratings

$$\mathbf{X}_i = \left(X_{i1}, X_{i2}, \ldots, X_{ip}\right)^T, \quad i = 1, 2, \ldots, n_1, \tag{3}$$

and n_2 diseased participants with ratings

$$\mathbf{Y}_i = \left(Y_{i1}, Y_{i2}, \ldots, Y_{ip}\right)^T, \quad i = 1, 2, \ldots, n_2. \tag{4}$$

The best linear combination coefficient $\mathbf{c} = (c_1, c_2, \ldots, c_p)^T$ which maximizes the pAUC can be estimated by maximizing the empirical estimate of pAUC, i.e.,

$$\begin{aligned} &\widehat{pAUC} \\ &\quad = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\left(\mathbf{X}_i^T \hat{\mathbf{c}} < \mathbf{Y}_j^T \hat{\mathbf{c}} \textit{ and } \mathbf{X}_i^T \hat{\mathbf{c}} > Q\left(1 - t_0\right)\right) \end{aligned} \tag{5}$$

where $Q(1 - t_0)$ is the $(1 - t_0)$ quantile of the empirical distribution of $\mathbf{X}_\mathrm{i}^\mathrm{T}\hat{\mathbf{c}}$.

2.2. Su and Liu's Method for pAUC. Assume that $\mathbf{X}_i$ and $\mathbf{Y}_j$ follow multivariate normal distribution with mean vector $\boldsymbol{\mu}_x, \boldsymbol{\mu}_y \in \Re^p$ and $p \times p$ covariance matrices $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$, i.e., $\mathbf{X}_i \sim MVN(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and $\mathbf{Y}_j \sim MVN(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$, respectively. Su and Liu derived the best linear combination coefficient $\mathbf{c} = (c_1, c_2, \ldots, c_p)^T$ that can maximize AUC based on the invariance property of ROC curve to scalar transformation and Fisher's discriminant coefficient [11]. When the two covariance matrices are equal or proportional to each other, the best linear coefficient based on Su and Liu's method also generates the ROC curve dominating all the others within any range of specificities.

2.3. Liu et al.'s Method for pAUC. By realizing the unsatisfactory performance from the use of Su and Liu's best linear combination coefficient, Liu et al. considered the scenario where $\mathbf{\Sigma}_x \neq \mathbf{\Sigma}_y$ [12]. The authors provided an approach to construct best linear combination that can maximize sensitivity over a certain range of specificities. In particular, if the high specificity region of an ROC curve is of interest, then the best linear combination coefficient is proportional to

$$\mathbf{\Sigma}_y^{-1/2}\alpha_p \tag{6}$$

where α_p is the eigenvector corresponding to the smallest eigenvalue of matrix $\mathbf{\Sigma}_y^{-1/2}\mathbf{\Sigma}_x\mathbf{\Sigma}_y^{-1/2}$. It has been showed that this linear combination produces the ROC curve dominating any other ROC curves in some particular specificity ranges.

2.4. Logistic Regression for pAUC. The logistic regression has been widely used to predict binary outcomes by considering linear combination of multiple predictors [13]. It models the probability of disease for a given subject with covariates $\mathbf{X}_i$ by using the logit link function, i.e.,

$$\Pr\left(D_i = 1\right) = \frac{\exp\left(\beta_0 + \mathbf{X}_i^T\mathbf{c}\right)}{1 + \exp\left(\beta_0 + \mathbf{X}_i^T\mathbf{c}\right)}, \tag{7}$$

where β_0 is the intercept and $\mathbf{X}_i$ and $\mathbf{c}$ are defined as before. Under the assumption of generalized linear model, the estimate of $\mathbf{c}$ followed by the logistic regression can maximize the likelihood function of binary outcomes. Jin and Lu proved that this estimate also provides the highest sensitivity uniformly over the entire range of specificity. This implies that the best linear combination equals $\widehat{\mathbf{c}}$ resulting in an ROC curve which not only has the maximum full AUC, but also dominates any other ROC curves within any range of potential interest and therefore leads to the maximum pAUC.

2.5. Pepe and Thompon's Method for pAUC. Without distributional assumptions on the data $\mathbf{X}$ and $\mathbf{Y}$, Pepe and Thompson [16] considered maximizing AUC and pAUC through rank-based estimate, i.e., the Mann-Whitney U statistics [1]. For simplicity, they proposed an algorithm to search for optimal linear combinations with number of biomarkers equal to 2 (p=2), i.e., $\mathbf{X}_i = (X_{i1}, X_{i2})^T$ for $i = 1, 2, \ldots, n_1$ and $\mathbf{Y}_i = (Y_{i1}, Y_{i2})^T$ for $i = 1, 2, \ldots, n_2$. Based on the fact that the ROC curve is variant to scale transformation, in order to maximize AUC or pAUC, finding the best combination coefficient $\mathbf{c} = (c_1, c_2)^T$, where $c_1, c_2 \in (-\infty, +\infty)$ is equivalent to finding $\mathbf{c} = (1, \alpha)^T$, where $\alpha \in (-\infty, +\infty)$. Let $[0, fpf_0]$ denote the range of false positive of potential interest. The estimate of AUC based on the Mann-Whitney U statistics and the estimate of pAUC can be obtained as

$$A\widehat{U}C(\alpha) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I\left(X_{i1} + \alpha X_{i2} < Y_{j1} + \alpha Y_{j2}\right) \tag{8}$$

and

$$\begin{aligned} pA\widehat{U}C(\alpha) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I\Big(X_{i1} + \alpha X_{i2} < Y_{j1} \\ + \alpha Y_{j2} \textit{ and } X_{i1} + \alpha X_{i2} > Q\left(1 - fpf_0, \alpha\right)\Big), \end{aligned} \tag{9}$$

respectively, where $Q(1 - fpf_0, \alpha)$ is the $(1 - fpf_0)$ quantile of $X_{i1} + \alpha X_{i2}$. The authors chose to implement a semiparametric method based on Heagerty and Pepe [21] to estimate $Q(1 - fpf_0, \alpha)$, while they also pointed out that other quantile estimation methods may be applied.

2.6. Min-Max Method for AUC. Liu et al. considered the min-max combination of biomarkers [20]. Let

$$\begin{aligned} X_{i,\max} &= \max_{1\le k\le p} X_{ik}, \\ Y_{i,\max} &= \max_{1\le k\le p} Y_{ik} \end{aligned} \tag{10}$$

be the maximum value of p biomarkers for nondiseased and diseased participants, respectively. Similarly, let

$$\begin{aligned} X_{i,\min} &= \min_{1\le k\le p} X_{ik}, \\ Y_{i,\min} &= \min_{1\le k\le p} Y_{ik} \end{aligned} \tag{11}$$

be the minimum value of p biomarkers for nondiseased and diseased participants, respectively.

The nonparametric estimate of AUC based on the Mann-Whitney U statistics by linearly combining the minimum and maximum values of p biomarkers for each subject can be obtained as

$$\begin{aligned} &A\widehat{U}C(\alpha) \\ &= \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I\left(X_{i,\max} + \alpha X_{i,\min} < Y_{j,\max} + \alpha Y_{j,\min}\right). \end{aligned} \tag{12}$$

Since this is not a continuous function of α, a search rather than a derivative-based method is required for the maximization. The searching method for the best value of α is exactly the same as Pepe and Thompson's method.

3. Methodology Extension: Min-Max Method for pAUC

We extend the min-max method to maximize the pAUC. Let $[0, fpf_0]$ denote the range of false positive of potential interest. By considering only the minimum and maximum values of p biomarkers for each individual, it follows that the nonparametric estimate of pAUC can be obtained as

$$\begin{aligned} \widehat{pAUC}(\alpha) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I\Big(X_{i,\max} + \alpha X_{i,\min} < Y_{j,\max} \\ + \alpha Y_{j,\min} \textit{ and } X_{i,\max} + \alpha X_{i,\min} > Q\left(1 - fpf_0, \alpha\right)\Big) \end{aligned} \tag{13}$$

where $Q(1-fpf_0,\alpha)$ is the $(1-fpf_0)$ quantile of $X_{i,\max}+\alpha X_{i,\min}$. For simplicity, the $(1-fpf_0)$ quantile of the empirical distribution of $X_{i,\max}+\alpha X_{i,\min}$ can be used to estimate $Q(1-fpf_0,\alpha)$. Then the Pepe and Thompson's [16] algorithm can be applied to search for the optimal value of α to maximize the estimate of the pAUC.

The new marker $(X_{i,\max}, Y_{i,\max})$ has larger sensitivity and smaller specificity for any given threshold c than any other individual marker, given that

$$\begin{aligned}\Pr\{Y_{i,\max} > c\} &= 1-\Pr\{Y_{i,\max}\le c\}\\ &= 1-\Pr\{Y_{ij}\le c, 1\le j\le p\}\\ &\ge 1-\Pr\{Y_{ik}\le c\} = \Pr\{Y_{ik}>c\}\end{aligned} \tag{14}$$

and

$$\begin{aligned}\Pr\{X_{i,\max}\le c\} &= \Pr\{X_{ij}\le c, 1\le j\le p\}\\ &\le \Pr\{X_{ik}\le c\}\end{aligned} \tag{15}$$

for all $1\le k\le p$; similarly, the new marker $(X_{i,\min}, Y_{i,\min})$ has smaller sensitivity and larger specificity for any given threshold c than any other individual marker, given that

$$\begin{aligned}\Pr\{Y_{i,\min} > c\} &= \Pr\{Y_{ij} > c,\ 1\le j\le p\}\\ &\le \Pr\{Y_{ik} > c\}\end{aligned} \tag{16}$$

and

$$\begin{aligned}\Pr\{X_{i,\min}\le c\} &= 1-\Pr\{X_{i,\min} > c\}\\ &= 1-\Pr\{X_{ij} > c,\ 1\le j\le p\}\\ &\ge 1-\Pr\{X_{ik} > c\} = \Pr\{X_{ik}\le c\}\end{aligned} \tag{17}$$

for all $1\le k\le p$. Therefore, we expect that the linear combination of the min-max biomarkers may provide larger partial area under the ROC curve than other methods. We employ simulation study to investigate how well the proposed method performs compared to other established methods.

The cross-validation has been widely used to evaluate the generalizability of the statistical results. Huang et al. [22] proposed a LOPO approach to evaluating the performance of the linear combination coefficient to estimate AUC for future observations. The estimate of AUC based on LOPO cross-validation is as follows:

$$\widehat{pAUC}^{CV} = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I\left(\mathbf{X}_i^T\widehat{\mathbf{c}}^{(-ij)} < \mathbf{Y}_j^T\widehat{\mathbf{c}}^{(-ij)}\right) \tag{18}$$

where $\widehat{\mathbf{c}}^{(-ij)}$ is the best linear combination coefficient based on the observed data without both the ith observation from nondiseased subject and the jth observation from diseased subject. They also demonstrated that the 5-fold and 10-fold cross-validation can be computationally efficient and the resulting estimate can be asymptotically unbiased for the future observations.

We implement the LOPO cross-validation on the pAUC to evaluate the generalizability of the statistical results. The estimate of the pAUC based on the LOPO cross-validation can be obtained as

$$\begin{aligned}\widehat{pAUC}^{CV} &= \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I\left(\mathbf{X}_i^T\widehat{\mathbf{c}}^{(-ij)}\right.\\ &\quad \left.< \mathbf{Y}_j^T\widehat{\mathbf{c}}^{(-ij)}\ and\ \mathbf{X}_i^T\widehat{\mathbf{c}}^{(-ij)} > Q\left(1-fpf_0,\alpha\right)\right)\end{aligned} \tag{19}$$

where $Q(1-fpf_0,\alpha)$ is the $(1-fpf_0)$ quantile of $\mathbf{X}_i^T\widehat{\mathbf{c}}^{(-ij)}$. For simplicity, in our simulation study the $(1-fpf_0)$ quantile of the empirical distribution of $\mathbf{X}_i^T\widehat{\mathbf{c}}^{(-ij)}$ will be used to estimate of $Q(1-fpf_0,\alpha)$.

4. Simulation

4.1. Description of Simulations. We conducted extensive simulation studies to investigate the performance of our proposed method with established combination methods based on the partial area under the ROC curves. Ratings of participants were simulated from different multivariate distributions with equal and unequal variance-covariance matrices. We examined false positive fraction ranges 0 – 0.1 and 0 – 0.2 and we considered different samples sizes: 50:50, 50:100, 100:50, and 100:100 for nondiseased and diseased participants, respectively.

For each simulated dataset, we computed the pAUC based on four different approaches: (1) min-max, denoted as MIN-MAX; (2) Su and Liu's [11], denoted as SULIU; (3) Liu et al.'s (2006), denoted as LIU; and the (4) logistic regression, denoted as LOGISTIC. In addition, we utilized two estimation methods: the re-substitution (denoted as Re-Sub) and 10-fold leave-one-pair-out cross-validation (denoted as LOPO) in computing the pAUC. The re-substitution method estimated the pAUC based on the linear combination of the coefficients derived using all the data for each method. The re-substitution method is usually overoptimistic for estimating the diagnostic/prognostic accuracy for future observations due to the reason between training set and validation set in the discipline of machine learning [22]. We obtained the mean of the pAUC by averaging over the 1,000 simulations, and standard deviation was the square root of the estimated sample variance of the estimated pAUC from 1,000 simulated datasets.

4.2. Multivariate Normal Distributions with Equal Variance-Covariance. We first compared the performance of the min-max approach on the pAUC with the other methods by generating dataset consisting of ratings from multivariate normal distributions (p=4) with different mean vectors and equal variance-covariance matrices (scenario #1). Exploiting the invariance property of the ROC curve to monotonically increasing transformation of the ratings, the distributions of ratings of nondiseased participants were set to be a multivariate normal distribution with mean $\boldsymbol{\mu}_x = (0,0,0,0)^T$ and variance-covariance matrix

$$\boldsymbol{\Sigma}_x = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}. \tag{20}$$

Under this scenario, ratings of diseased participants were generated from multivariate normal distributions with variance-covariance matrix $\boldsymbol{\Sigma}_y$ equal to $\boldsymbol{\Sigma}_x$, and the mean vectors were selected to generate the AUC equal to 0.70, 0.73, 0.76, and 0.80 for markers # 1, 2, 3, and 4, respectively (Case #1), and the AUC equal to 0.6, 0.7, 0.8, and 0.9 for markers # 1, 2, 3, and 4, respectively (Case #2).

4.3. Multivariate Normal Distributions with Unequal Variance-Covariance. We also considered multivariate normal distributions with different mean and unequal variance-covariance matrices for nondiseased and diseased participants (scenario #2). The mean settings are the same as Case 1 and Case 2 as discussed in scenario 1. The variance-covariance matrices were

$$\boldsymbol{\Sigma}_x = \begin{pmatrix} 1 & 0.2 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 & 0.2 \\ 0.2 & 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 0.2 & 1 \end{pmatrix}, \quad \text{and } \boldsymbol{\Sigma}_y = \begin{pmatrix} 1 & 0.8 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 & 0.8 \\ 0.8 & 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 0.8 & 1 \end{pmatrix}. \tag{21}$$

4.4. Multivariate Log-Normal Distributions with Unequal Variance-Covariance. We investigated the performance of the different combination methods by generating dataset consisting of ratings from multivariate log-normal distributions (scenario #3). Ratings were first generated similarly to scenario #2 and then exponentiated to obtain the multivariate log-normal marker values.

4.5. Multivariate Gamma Distributions. We further examined the performance of the different combination methods by generating gamma ROC curves with the AUC settings in Case 1 and Case 2 (scenario #4). The gamma family is one of the well-known families of ROC curves [9, 10, 23–26]. Due to the concavity and flexibility in the shape, Ma et al. [9] and Ma et al. [10] demonstrated that the families of gamma ROC curves provided practically reasonable straight-line shaped concave ROC curves, where the statistical inference based on pAUCs is preferable.

The probability density function of the underlying rating model of the gamma ROC curve has the following form:

$$f(x;\kappa,\theta) = \frac{1}{\theta^{\kappa}}\frac{1}{\tau(\kappa)}x^{\kappa-1}e^{-x/\theta}. \tag{22}$$

When κ approaches 0, the gamma ROC curve approaches the shape of a straight-line and when $\kappa > 1$ the shape of the gamma ROC curve resembles an ROC curve with latent normality assumptions. When κ=1 the gamma ROC curve is equivalent to the power-law ROC curve [23, 27]. Here we are interested in the investigation of a scenario with straight-line shaped gamma ROC curves (κ=1/3), because this type of ROC curves cannot be generated by the previous scenarios.

Each simulated dataset consisted of ratings generated from multivariate gamma distributions with κ=1/3. Due to the invariance property of the ROC curves, without any loss of generality, we set θ=1 for latent ratings of nondiseased participants. We then selected θ for the latent diseased ratings to reflect the targeted area under the ROC curve in Case #1 and Case #2. The between-modality correlation of 0.5 was established using a Gaussian copula model [28]. All the programs were written by the first author in R version 2.15.3 and are available: https://duke.box.com/s/u32h7aayxd9bjo41b619xpb21sj1nm67.

4.6. Simulation Results. We compared the performance of the min-max method in estimating the pAUC with three established methods assuming the ratings are from multivariate normal distributions with equal variance-covariance matrices (Table 1). The SULIU and LOGISTIC almost always performed better than the min-max and LIU based on the pAUCs estimated from both the re-substitution and the LOPO cross-validation. In addition, the performances of SULIU and LOGISTIC approaches were similar when the AUCs were either close or further apart. The min-max approach produced slightly smaller pAUC estimates than that of SULIU and LOGISTIC when the AUCs among biomarkers were relatively close (i.e., Case #1), while this approach became worse when the AUCs were far apart (i.e., Case #2).

Moreover, we examined the performance of the four methods, i.e., MIN-MAX, SULIU, LIU, and LOGISTIC, assuming ratings are from multivariate normal distributions with unequal variance-covariance matrices (Table 2). When the AUCs were close (Case #1), the min-max method was superior to the other methods in terms of its ability to produce the largest pAUCs based on both the re-substitution and the LOPO cross-validation. When the AUCs were far apart (i.e., Case #2), the SULIU and LOGISTIC methods had similar performances superior to the other two methods. The SULIU method was slightly better than the LOGISTIC based on the LOPO cross-validation since this takes into account the normality of data with unequal variance-covariance matrices. It should be noted that the difference in the estimates of the pAUCs between the re-substitution and the LOPO cross-validation was very small under this scenario.

Furthermore, we studied the performance of the different combination methods assuming multivariate log-normal distributions. From Table 3, under this scenario where data are highly skewed, the min-max approach dominated the other approaches when the AUCs were close (Case #1). On the other hand, the LOGISTIC approach performed better when the AUCs are far apart. It is interesting to observe that the LIU method was suboptimal under both cases in terms of its ability to estimate the pAUC through the LOPO cross-validation whereas the SULIU method had

Table 1: Means (standard deviations) of the partial area under the ROC curve for different combination methods based on the dataset consisted of ratings from multivariate normal distributions with equal variance-covariance matrices (scenario#1) with 1000 simulated datasets.

AUCs	Fpf range	Sample Size		MIN-MAX	SULIU	LIU	LOGISTIC
0.7-0.8	0-0.1	50:50	Re-Sub	0.036 (0.010)	0.038 (0.011)	0.024 (0.011)	0.038 (0.011)
			LOPO	0.028 (0.012)	0.030 (0.011)	0.016 (0.011)	0.030 (0.011)
		50:100	Re-Sub	0.036 (0.009)	0.038 (0.010)	0.024 (0.011)	0.037 (0.010)
			LOPO	0.028 (0.011)	0.031 (0.010)	0.016 (0.011)	0.031 (0.010)
		100:50	Re-Sub	0.034 (0.008)	0.036 (0.009)	0.021 (0.010)	0.036 (0.009)
			LOPO	0.028 (0.009)	0.031 (0.009)	0.015 (0.009)	0.030 (0.009)
		100:100	Re-Sub	0.033 (0.007)	0.036 (0.007)	0.021 (0.010)	0.036 (0.007)
			LOPO	0.028 (0.008)	0.031 (0.007)	0.015 (0.009)	0.031 (0.007)
	0-0.2	50:50	Re-Sub	0.094 (0.018)	0.101 (0.020)	0.064 (0.024)	0.101 (0.020)
			LOPO	0.081 (0.022)	0.086 (0.021)	0.049 (0.025)	0.086 (0.021)
		50:100	Re-Sub	0.092 (0.017)	0.099 (0.018)	0.064 (0.024)	0.099 (0.018)
			LOPO	0.081 (0.021)	0.087 (0.019)	0.049 (0.024)	0.087 (0.019)
		100:50	Re-Sub	0.091 (0.015)	0.097 (0.016)	0.059 (0.022)	0.098 (0.016)
			LOPO	0.082 (0.018)	0.087 (0.017)	0.047 (0.022)	0.087 (0.017)
		100:100	Re-Sub	0.089 (0.013)	0.096 (0.013)	0.058 (0.022)	0.096 (0.013)
			LOPO	0.081 (0.016)	0.089 (0.014)	0.047 (0.021)	0.089 (0.014)
0.6-0.9	0-0.1	50:50	Re-Sub	0.047 (0.010)	0.064 (0.011)	0.034 (0.018)	0.065 (0.011)
			LOPO	0.040 (0.012)	0.056 (0.013)	0.026 (0.017)	0.054 (0.012)
		50:100	Re-Sub	0.046 (0.010)	0.063 (0.010)	0.034 (0.018)	0.063 (0.010)
			LOPO	0.041 (0.012)	0.057 (0.012)	0.025 (0.016)	0.056 (0.011)
		100:50	Re-Sub	0.045 (0.009)	0.062 (0.009)	0.029 (0.016)	0.062 (0.009)
			LOPO	0.040 (0.010)	0.057 (0.010)	0.023 (0.014)	0.056 (0.010)
		100:100	Re-Sub	0.044 (0.008)	0.062 (0.008)	0.030 (0.016)	0.062 (0.008)
			LOPO	0.040 (0.009)	0.058 (0.008)	0.024 (0.014)	0.057 (0.008)
	0-0.2	50:50	Re-Sub	0.114 (0.017)	0.148 (0.017)	0.084 (0.037)	0.149 (0.017)
			LOPO	0.106 (0.020)	0.137 (0.019)	0.070 (0.036)	0.135 (0.019)
		50:100	Re-Sub	0.114 (0.017)	0.147 (0.015)	0.085 (0.037)	0.147 (0.015)
			LOPO	0.107 (0.019)	0.138 (0.017)	0.070 (0.034)	0.136 (0.017)
		100:50	Re-Sub	0.112 (0.015)	0.146 (0.014)	0.075 (0.033)	0.146 (0.014)
			LOPO	0.106 (0.017)	0.138 (0.015)	0.065 (0.032)	0.137 (0.015)
		100:100	Re-Sub	0.110 (0.013)	0.146 (0.012)	0.079 (0.035)	0.146 (0.012)
			LOPO	0.106 (0.014)	0.140 (0.012)	0.068 (0.032)	0.139 (0.012)

Table 2: Means and standard deviation (SD) of the partial area under the ROC curve for different combination methods based on the dataset consisted of ratings from multivariate normal distributions with unequal variance-covariance matrices (scenario#2) with 1000 simulated datasets.

AUCs	Fpf range	Sample Size		MIN-MAX	SULIU	LIU	LOGISTIC
0.7-0.8	0-0.1	50:50	Re-Sub	0.059 (0.011)	0.044 (0.011)	0.046 (0.010)	0.044 (0.011)
			LOPO	0.052 (0.013)	0.035 (0.012)	0.042 (0.010)	0.034 (0.012)
		50:100	Re-Sub	0.058 (0.009)	0.044 (0.009)	0.046 (0.008)	0.042 (0.010)
			LOPO	0.053 (0.012)	0.036 (0.010)	0.042 (0.009)	0.033 (0.010)
		100:50	Re-Sub	0.057 (0.008)	0.043 (0.009)	0.045 (0.008)	0.045 (0.009)
			LOPO	0.052 (0.010)	0.037 (0.009)	0.043 (0.008)	0.039 (0.009)
		100:100	Re-Sub	0.057 (0.007)	0.043 (0.008)	0.044 (0.007)	0.043 (0.008)
			LOPO	0.053 (0.009)	0.038 (0.008)	0.042 (0.007)	0.038 (0.008)
	0-0.2	50:50	Re-Sub	0.136 (0.018)	0.109 (0.019)	0.109 (0.018)	0.109 (0.019)
			LOPO	0.128 (0.021)	0.093 (0.021)	0.102 (0.018)	0.093 (0.021)
		50:100	Re-Sub	0.135 (0.015)	0.109 (0.016)	0.109 (0.015)	0.106 (0.017)
			LOPO	0.129 (0.018)	0.095 (0.018)	0.103 (0.016)	0.090 (0.019)
		100:50	Re-Sub	0.133 (0.014)	0.107 (0.016)	0.107 (0.015)	0.110 (0.015)
			LOPO	0.128 (0.016)	0.098 (0.017)	0.104 (0.015)	0.101 (0.016)
		100:100	Re-Sub	0.133 (0.012)	0.106 (0.013)	0.107 (0.012)	0.107 (0.013)
			LOPO	0.129 (0.013)	0.099 (0.014)	0.104 (0.012)	0.099 (0.014)
0.6-0.9	0-0.1	50:50	Re-Sub	0.051 (0.010)	0.058 (0.012)	0.049 (0.011)	0.059 (0.013)
			LOPO	0.044 (0.012)	0.048 (0.014)	0.045 (0.012)	0.045 (0.014)
		50:100	Re-Sub	0.050 (0.008)	0.059 (0.012)	0.049 (0.010)	0.056 (0.013)
			LOPO	0.044 (0.010)	0.049 (0.013)	0.046 (0.010)	0.045 (0.013)
		100:50	Re-Sub	0.049 (0.008)	0.057 (0.010)	0.048 (0.009)	0.059 (0.010)
			LOPO	0.044 (0.009)	0.051 (0.011)	0.046 (0.009)	0.051 (0.010)
		100:100	Re-Sub	0.049 (0.007)	0.057 (0.009)	0.048 (0.008)	0.056 (0.009)
			LOPO	0.044 (0.008)	0.051 (0.010)	0.046 (0.008)	0.049 (0.010)
	0-0.2	50:50	Re-Sub	0.118 (0.017)	0.143 (0.018)	0.114 (0.020)	0.143 (0.019)
			LOPO	0.108 (0.020)	0.128 (0.021)	0.108 (0.021)	0.124 (0.021)
		50:100	Re-Sub	0.117 (0.014)	0.143 (0.017)	0.115 (0.018)	0.141 (0.018)
			LOPO	0.109 (0.016)	0.129 (0.020)	0.109 (0.019)	0.124 (0.020)
		100:50	Re-Sub	0.116 (0.014)	0.141 (0.016)	0.113 (0.016)	0.143 (0.015)
			LOPO	0.109 (0.016)	0.133 (0.017)	0.110 (0.017)	0.132 (0.016)
		100:100	Re-Sub	0.115 (0.012)	0.140 (0.013)	0.113 (0.014)	0.140 (0.013)
			LOPO	0.109 (0.013)	0.133 (0.014)	0.110 (0.014)	0.131 (0.014)

Table 3: Means and standard deviation (SD) of the partial area under the ROC curve for different combination methods based on the dataset consisted of ratings from multivariate log-normal distributions with unequal variance-covariance matrices (scenario#3) with 1,000 simulated datasets.

AUCs	Fpf range	Sample Size		MIN-MAX	SULIU	LIU	LOGISTIC
0.7-0.8	0-0.1	50:50	Re-Sub	0.059 (0.011)	0.035 (0.009)	0.040 (0.010)	0.040 (0.010)
			LOPO	0.054 (0.012)	0.026 (0.010)	0.031 (0.011)	0.028 (0.011)
		50:100	Re-Sub	0.058 (0.009)	0.035 (0.008)	0.040 (0.009)	0.039 (0.009)
			LOPO	0.054 (0.011)	0.028 (0.009)	0.032 (0.010)	0.028 (0.010)
		100:50	Re-Sub	0.057 (0.008)	0.033 (0.008)	0.037 (0.008)	0.038 (0.008)
			LOPO	0.054 (0.009)	0.027 (0.008)	0.031 (0.009)	0.030 (0.009)
		100:100	Re-Sub	0.057 (0.007)	0.033 (0.007)	0.036 (0.007)	0.037 (0.007)
			LOPO	0.054 (0.008)	0.028 (0.007)	0.031 (0.008)	0.030 (0.007)
	0-0.2	50:50	Re-Sub	0.136 (0.018)	0.090 (0.018)	0.095 (0.019)	0.099 (0.019)
			LOPO	0.129 (0.020)	0.074 (0.020)	0.082 (0.020)	0.079 (0.020)
		50:100	Re-Sub	0.135 (0.015)	0.091 (0.015)	0.096 (0.017)	0.097 (0.016)
			LOPO	0.129 (0.017)	0.077 (0.017)	0.084 (0.018)	0.079 (0.018)
		100:50	Re-Sub	0.133 (0.014)	0.088 (0.015)	0.091 (0.016)	0.095 (0.016)
			LOPO	0.129 (0.015)	0.077 (0.017)	0.084 (0.016)	0.084 (0.017)
		100:100	Re-Sub	0.133 (0.012)	0.087 (0.013)	0.091 (0.014)	0.094 (0.013)
			LOPO	0.130 (0.013)	0.079 (0.014)	0.084 (0.014)	0.083 (0.014)
0.6-0.9	0-0.1	50:50	Re-Sub	0.050 (0.010)	0.051 (0.012)	0.056 (0.012)	0.059 (0.012)
			LOPO	0.043 (0.012)	0.043 (0.013)	0.048 (0.014)	0.048 (0.013)
		50:100	Re-Sub	0.049 (0.008)	0.050 (0.011)	0.057 (0.011)	0.058 (0.011)
			LOPO	0.043 (0.011)	0.044 (0.012)	0.050 (0.013)	0.050 (0.012)
		100:50	Re-Sub	0.114 (0.014)	0.123 (0.020)	0.128 (0.018)	0.138 (0.016)
			LOPO	0.107 (0.016)	0.115 (0.021)	0.122 (0.019)	0.129 (0.017)
		100:100	Re-Sub	0.047 (0.007)	0.049 (0.010)	0.054 (0.009)	0.056 (0.009)
			LOPO	0.044 (0.008)	0.045 (0.010)	0.050 (0.010)	0.051 (0.009)
	0-0.2	50:50	Re-Sub	0.116 (0.017)	0.126 (0.021)	0.130 (0.022)	0.141 (0.019)
			LOPO	0.106 (0.021)	0.114 (0.023)	0.118 (0.024)	0.125 (0.021)
		50:100	Re-Sub	0.115 (0.014)	0.125 (0.020)	0.131 (0.020)	0.140 (0.018)
			LOPO	0.107 (0.017)	0.114 (0.021)	0.121 (0.022)	0.128 (0.019)
		100:50	Re-Sub	0.048 (0.008)	0.049 (0.011)	0.054 (0.010)	0.057 (0.010)
			LOPO	0.044 (0.009)	0.044 (0.011)	0.050 (0.011)	0.051 (0.010)
		100:100	Re-Sub	0.113 (0.012)	0.124 (0.017)	0.128 (0.016)	0.138 (0.014)
			LOPO	0.108 (0.014)	0.117 (0.018)	0.122 (0.017)	0.130 (0.015)

the worst performance since the normality assumption was violated.

Lastly, we considered the performance of different combination methods by generating gamma ROC curves. From Table 4, (Scenario #4) where data suggest a straight-line shape ROC curve, when the AUCs were close, the min-max approach performed better than the other three approaches in obtaining the largest pAUCs through both the re-substitution and the LOPO cross-validation. When the AUCs were far apart (Case #2), the min-max approach yielded the best pAUC estimates through LOPO cross-validation. The LOGITIC approach was best based on the re-substitution.

5. Example

5.1. Example 1. We used data from Cancer and Leukemia Group B study 90206, a Phase III clinical trial of metastatic renal-cell carcinoma [29, 30], to provide an example of our proposed method. The study randomized 732 patients, 369 to anti-VEGF treatment and 363 to a control group [29, 30]. The trial was designed with 588 deaths so that the log-rank statistic would have 86% power to detect a hazard ratio of 0.76 for deaths assuming a two-sided significance level of 0.05. The trial collected plasma from patients in order to study the relationship of angiogenic and inflammatory markers with clinical outcomes. A primary objective of the correlative science study was to associate the anti-VEGF biomarkers from the angioma assay with overall survival and build a prognostic model that predicts the clinical outcome [31, 32]. Another objective was to correlate the anti-VEGF biomarkers with the best objective response rate (defined as either partial or complete response). The angioma multiplex array has gone through a rigorous evaluation to ensure data quality [31, 32]. Markers performed include Ang-2, bFGF, BMP-9, CRP, Endoglin, Gro-a, HGF, ICAM-1, IGFBP-1, IGFBP-2, IGFBP-3, IL-6, IL-8, MCP-1, OPN, P-selectin, Pai-1-active, Pai-1-total, PDGF-AA, PDGF-BB, PEDF, PlGF, SDF-1, TGFβ1, TGFβ2, TGFβ3-R3, TSP-2, VCAM-1, VEGF, VEGF-C, VEGF-D, VEGF-R1, and VEGF-R2.

We used the random forest, LASSO, and adaptive LASSO to select the top three biomarkers of the 33 biomarkers for best objective response. The top three genes (HGF, IL_6, and VEGF_R2) with highest full AUC (0.576, 0.610, and 0.563) were chosen as an example to demonstrate the scenario where the AUCs were close to each other as a potential advantage of the use of the proposed method. The empirical estimates for the pAUC for these three biomarkers are 0.012, 0.012, and 0.028. The correlation matrices for nonresponders and responders are

$$Corr_x = \begin{pmatrix} 1 & 0.530 & 0.319 \\ 0.530 & 1 & 0.273 \\ 0.319 & 0.273 & 1 \end{pmatrix}$$
$$\text{and } Corr_y = \begin{pmatrix} 1 & 0.453 & 0.219 \\ 0.453 & 1 & 0.225 \\ 0.219 & 0.225 & 1 \end{pmatrix}. \tag{23}$$

The proposed method provided the following combination:

$$\max\{HGF, IL\_6,\ and\ VEGF\_R2\} - \min\{HGF, IL\_6,\ and\ VEGF\_R2\} \tag{24}$$

with the estimated pAUC of 0.0427 and the estimated standard deviation of 0.0080 based on 1,000 bootstrap sampling.

In contrast, the SULIU method provided the following combination:

$$HGF + 1.36 * IL\_6 - 1.81 * VEGF\_R2 \tag{25}$$

with the estimated pAUC of 0.0426 and the estimated standard deviation of 0.0084.

The LIU method provided the following combination:

$$HGF - 1.21 * IL\_6 - 0.06 * VEGF\_R2 \tag{26}$$

with the estimated pAUC of 0.0254 and the estimated standard deviation of 0.0099, whereas the LOGISTIC's method had the following combination:

$$HGF + 1.52 * IL\_6 - 1.88 * VEGF\_R2 \tag{27}$$

with the estimated pAUC of 0.0422 and the estimated standard deviation of 0.0084.

5.2. Example 2. In this section, the proposed method MIN-MAX as well as the SULIU, LIU, and the LOGISTIC is applied to a real dataset of 125 females on Duchenne Muscular Dystrophy (DMD) dataset. This biomedical data originally containing 209 observations (134 for "normals" and 75 for "carriers") has been studied by Cox et al. [33] in order to develop screening methods to identify carriers of a rare genetic disorder based on four measurements made on blood samples. This dataset has been widely studied in the literature for improving the classification accuracy by using ROC analysis. The main objective is to combine four markers to increase the diagnostic accuracy of screening females as potential DMD carriers. For example, Kang et al. [14] applied the stepwise methods to combine four makers in this data to improve AUC; Hsu and Hsueh [18] and Yu and Park [19] applied their proposed algorithm to pAUC in this data.

Since four different variables M1–M4 were measured in each blood sample, we processed the data by taking average values for each measurement if one had blood drawn at several different time points. Among the 125 females, there are 87 normals and 38 carriers.

Similarly, we investigated the performance of the four different methods on the pAUC over the range 0–0.2. Since the four measurements are in different scales, we applied the standardization method by dividing each value by the range of that variable before the use of MIN-MAX approach. $M1^*$- $M4^*$ denote the standardized marker values. The empirical estimates for the pAUC for these four biomarkers are 0.1472, 0.0436, 0.1086, and 0.1229 for the M1–M4, respectively. The empirical estimates for the full AUC are 0.9034, 0.6057, 0.8232, and 0.8814. The correlation matrices for nonrespondents and respondents are

Table 4: Means and standard deviation (SD) of the partial area under the ROC curve for different combination methods based on the dataset consisted of ratings from multivariate gamma distributions (scenario#4) with 1000 simulated datasets.

AUCs	Fpf range	Sample Size		MIN-MAX	SULIU	LIU	LOGISTIC
0.7-0.8	0-0.1	50:50	Re-Sub	0.068 (0.007)	0.059 (0.008)	0.052 (0.010)	0.066 (0.008)
			LOPO	0.067 (0.008)	0.056 (0.009)	0.048 (0.011)	0.059 (0.008)
		50:100	Re-Sub	0.068 (0.006)	0.060 (0.007)	0.052 (0.009)	0.065 (0.006)
			LOPO	0.067 (0.006)	0.058 (0.007)	0.049 (0.009)	0.060 (0.007)
		100:50	Re-Sub	0.068 (0.007)	0.059 (0.008)	0.050 (0.009)	0.065 (0.007)
			LOPO	0.067 (0.007)	0.056 (0.008)	0.048 (0.010)	0.060 (0.007)
		100:100	Re-Sub	0.068 (0.005)	0.059 (0.006)	0.051 (0.007)	0.064 (0.005)
			LOPO	0.067 (0.005)	0.057 (0.006)	0.049 (0.008)	0.061 (0.006)
	0-0.2	50:50	Re-Sub	0.145 (0.013)	0.130 (0.015)	0.111 (0.020)	0.141 (0.014)
			LOPO	0.143 (0.014)	0.125 (0.016)	0.105 (0.022)	0.131 (0.015)
		50:100	Re-Sub	0.146 (0.010)	0.132 (0.012)	0.111 (0.017)	0.141 (0.011)
			LOPO	0.144 (0.011)	0.129 (0.013)	0.107 (0.018)	0.133 (0.012)
		100:50	Re-Sub	0.145 (0.012)	0.130 (0.014)	0.109 (0.018)	0.140 (0.013)
			LOPO	0.144 (0.012)	0.125 (0.015)	0.106 (0.020)	0.133 (0.013)
		100:100	Re-Sub	0.145 (0.009)	0.131 (0.011)	0.110 (0.015)	0.139 (0.010)
			LOPO	0.144 (0.009)	0.128 (0.011)	0.108 (0.015)	0.134 (0.010)
0.6-0.9	0-0.1	50:50	Re-Sub	0.081 (0.006)	0.065 (0.009)	0.076 (0.007)	0.084 (0.006)
			LOPO	0.081 (0.006)	0.060 (0.011)	0.076 (0.007)	0.079 (0.008)
		50:100	Re-Sub	0.081 (0.004)	0.066 (0.008)	0.076 (0.005)	0.083 (0.004)
			LOPO	0.081 (0.004)	0.062 (0.009)	0.075 (0.005)	0.080 (0.005)
		100:50	Re-Sub	0.081 (0.005)	0.065 (0.009)	0.076 (0.006)	0.084 (0.005)
			LOPO	0.081 (0.005)	0.061 (0.010)	0.075 (0.006)	0.079 (0.008)
		100:100	Re-Sub	0.081 (0.004)	0.065 (0.008)	0.076 (0.005)	0.083 (0.004)
			LOPO	0.081 (0.004)	0.063 (0.008)	0.075 (0.005)	0.081 (0.004)
	0-0.2	50:50	Re-Sub	0.167 (0.010)	0.141 (0.016)	0.157 (0.013)	0.173 (0.010)
			LOPO	0.167 (0.011)	0.133 (0.018)	0.156 (0.013)	0.164 (0.013)
		50:100	Re-Sub	0.168 (0.008)	0.143 (0.013)	0.157 (0.010)	0.171 (0.008)
			LOPO	0.167 (0.008)	0.137 (0.014)	0.156 (0.010)	0.167 (0.008)
		100:50	Re-Sub	0.167 (0.010)	0.141 (0.016)	0.156 (0.012)	0.172 (0.010)
			LOPO	0.167 (0.010)	0.134 (0.017)	0.155 (0.012)	0.165 (0.013)
		100:100	Re-Sub	0.167 (0.007)	0.142 (0.013)	0.156 (0.009)	0.171 (0.007)
			LOPO	0.167 (0.007)	0.138 (0.013)	0.156 (0.009)	0.167 (0.007)

Table 5: The coefficients of the optimal linear combination and the corresponding estimated pAUC.

Method	M1	M2	M3	M4	pAUC
MIN-MAX	-	-	-	-	0.161
SULIU	1	12.6333	7.7165	13.6415	0.137
LIU	1	0.5248	0.7805	-0.1087	0.151
LOGISTIC	1	0.6950	1.3806	0.2545	0.156

$$Corr_x = \begin{pmatrix} 1 & -0.380 & 0.012 & 0.236 \\ -0.380 & 1 & 0.130 & 0.155 \\ 0.012 & 0.130 & 1 & 0.281 \\ 0.236 & 0.155 & 0.281 & 1 \end{pmatrix}$$

$$\text{and } Corr_y = \begin{pmatrix} 1 & -0.037 & 0.688 & 0.625 \\ -0.037 & 1 & -0.222 & -0.098 \\ 0.688 & -0.222 & 1 & 0.612 \\ 0.625 & -0.098 & 0.612 & 1 \end{pmatrix} \quad (28)$$

The proposed method provided the following combination (Table 5):

$$\max\{M1^*, M2^*, M3^*, M4\} + 6.6667 * \min\{M1^*, M2^*, M3^*, M4^*\} \quad (29)$$

with the estimated pAUC of 0.161 and the estimated standard deviation of 0.0119 based on 1,000 bootstrap sampling.

In contrast, the SULIU method provided the following combination (Table 5):

$$M1 + 12.6333 * M2 + 7.7165 * M3 + 13.6415 * M4 \quad (30)$$

with the estimated pAUC of 0.137 and the estimated standard deviation of 0.0157.

The LIU method provided the following combination (Table 5):

$$M1 + 0.5248 * M2 + 0.7805 * M3 - 0.1087 * M4 \quad (31)$$

with the estimated pAUC of 0.151 and the estimated standard deviation of 0.0135, whereas the LOGISTIC's method had the following combination (Table 5):

$$M1 + 0.6950 * M2 + 1.3806 * M3 + 0.2545 * M4 \quad (32)$$

with the estimated pAUC of 0.156 and the estimated standard deviation of 0.0138.

Figure 1 presents the performance for each method.

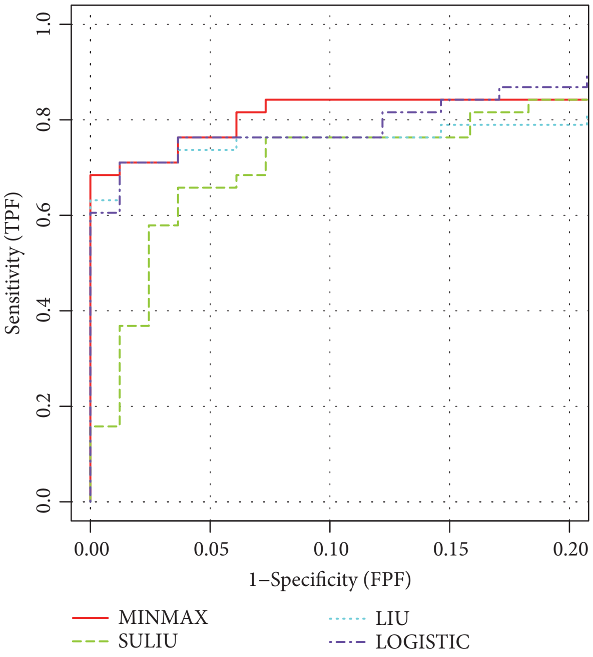


Figure 1

6. Discussion

In this article, we extend the min-max method to the estimation of the pAUC and compare its performances to three commonly utilized methods. The proposed method has the advantage of both the min-max method and Pepe and Thompson's method [16]. The expected advantages of this approach are threefold. First, it may yield larger partial area under the ROC curves. Second, it is a nonparametric approach and therefore it is robust against distributional assumptions. Lastly, it is computationally feasible and efficient since the min-max procedure involves searching for only one single coefficient. Our works [9, 10] have shown that the use of pAUC not only is clinically useful but also is statistically more efficient than the use of the full AUC in the families of area under the ROC curves that are nearly straight-line shaped. Another advantage of this method demonstrated through our simulation study is that in the scenario of straight-line shaped gamma ROC curves the estimate of pAUC based on re-substitution is close to the estimate based on the LOPO cross-validation. This implies that the min-max method on pAUC leads to good generalizability.

As pointed out by several authors [14, 22, 34], the use of the re-substitution to estimate the area under the ROC curve could usually lead to the overoptimistic result, or upward biased estimates for independent dataset, or future observations. Huang et al. [22] proposed to use the LOPO cross-validation to obtain less biased estimates. Kang et al. [14] applied the LOPO cross-validation to compare different combination methods to maximize the AUC. Because the estimates through cross-validation lead to more reliable results in terms of its ability to generalize to an independent dataset, we recommend using cross-validation which performs better when decisions based on the re-substitution and the cross-validation approaches are different. Based on our simulation results, it is not surprising to observe that the standard deviation of the estimated pAUC decreased as the sample size increased and that the estimate of the pAUC based on the re-substitution approach was becoming closer to the estimate of the pAUC based on the LOPO cross-validation as the sample size increased.

Evaluation of diagnostic assays and prognostic performance of biomarkers will continue to remain an important research topic in several medical areas. This is especially true in oncology where diagnostic assays based on several combinations of biomarkers are developed and validated. For example, a 22-gene model was developed and validated to predict prostate cancer risk [35]. In addition, identifying predictive markers of clinical outcomes is a hot area of research as finding the optimal treatment to tailor patients is attractive not only to patients but also to physicians, insurance company, and society as a whole. Currently, several predictors or signatures of outcomes are being used to guide therapies in clinical trials [35]. For example OncotypeDx, a 21-gene expression signature, is being used to select treatment in patients with breast cancer based on the recurrence score [36]. Recognizing the fact that more predictors will continue to be applied in the clinic, it is critical that when a combination of biomarkers is developed this would result in the highest pAUC.

Based on our extensive simulations, our recommendations are the following:

(1) Use the SULIU or LOGISTIC approach to estimate the pAUC with approximately equal variance multivariate normal data regardless whether the AUCs among biomarkers are relatively close or far apart. The LIU's approach underestimated the pAUC approximately by 1/3. This is partly due to the instability of the eigenvector of the identity matrix, since LIU's approach involves the calculation of the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{\Sigma}_y^{-1/2}\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_y^{-1/2}$ which is an identity matrix under this scenario when $\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_y$, and the eigenvector corresponding to the smallest eigenvalue is not stable under small perturbation of the identity matrix [15].

(2) Utilize the min-max approach to estimate the pAUC with unequal variance multivariate normal data when the AUCs are relatively close and use the SULIU's approach when the AUCs are far apart.

(3) Employ the min-max approach to estimate the pAUC with highly skewed data when the AUCs are relatively close, but use the LOGISTIC method when the AUCs are far apart.

(4) Use the min-max approach to estimate the pAUC with straight-line shaped ROC curves regardless whether the AUCs are close or far apart.

In summary, the min-max approach seems to be robust and investigators are encouraged to use it in the estimation of the pAUC. It is simple to implement and is computationally feasible. In an era of personalized medicine, it is anticipated that the evaluation of diagnostic assays and the performance of the combination of biomarkers will remain an important area of research not only in diagnosing patients but also in treating patients with the disease.

Data Availability

The data from the simulation are available from the first author. The data from CALGB 90206 can be accessed through the Alliance in Clinical trials in Oncology.

Disclosure

The content of this article was presented at the 2016 Eastern North American Region Annual Meeting in Austin, TX.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was funded in part by the NIH R01CA155296, U01CA157703, the Prostate Cancer Foundation Challenge Award, and the United States Army Medical Research (Awards W81XWH-15-1-0467 and W81XWH-18-1-0278). Research of A. Liu was supported by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development Intramural Research Program.

References

[1] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[2] X.-H. Zhou, N. A. Obuchowski, and D. K. McClish, *Statistical Methods in Diagnostic Medicine*, Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], New York, NY, USA, 2002.

[3] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, vol. 28 of *Oxford Statistical Science Series*, Oxford University Press, Oxford, UK, 2003.

[4] D. K. McClish, "Analyzing a Portion of the ROC Curve," *Medical Decision Making*, vol. 9, no. 3, pp. 190–195, 1989.

[5] Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," *Radiology*, vol. 201, no. 3, pp. 745–750, 1996.

[6] L. E. Dodd and M. S. Pepe, "Partial AUC estimation and regression," *Biometrics: Journal of the International Biometric Society*, vol. 59, no. 3, pp. 614–623, 2003.

[7] Y. He and M. Escobar, "Nonparametric statistical inference method for partial areas under receiver operating characteristic curves, with application to genomic studies," *Statistics in Medicine*, vol. 27, no. 25, pp. 5291–5308, 2008.

[8] D. D. Zhang, X.-H. Zhou, D. H. Freeman Jr., and J. L. Freeman, "A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets," *Statistics in Medicine*, vol. 21, no. 5, pp. 701–715, 2002.

[9] H. Ma, A. I. Bandos, H. E. Rockette, and D. Gur, "On use of partial area under the ROC curve for evaluation of diagnostic performance," *Statistics in Medicine*, vol. 32, no. 20, pp. 3449–3458, 2013.

[10] H. Ma, A. I. Bandos, and D. Gur, "On the use of partial area under the ROC curve for comparison of two diagnostic tests," *Biometrical Journal*, vol. 57, no. 2, pp. 304–310, 2015.

[11] J. Q. Su and J. S. Liu, "Linear combinations of multiple diagnostic markers," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1350–1355, 1993.

[12] A. Liu, E. F. Schisterman, and Y. Zhu, "On linear combinations of biomarkers to improve diagnostic accuracy," *Statistics in Medicine*, vol. 24, no. 1, pp. 37–47, 2005.

[13] H. Jin and Y. Lu, "The optimal linear combination of multiple predictors under the generalized linear models," *Statistics & Probability Letters*, vol. 79, no. 22, pp. 2321–2327, 2009.

[14] L. Kang, A. Liu, and L. Tian, "Linear combination methods to improve diagnostic/prognostic accuracy on future observations," *Statistical Methods in Medical Research*, vol. 25, no. 4, pp. 1359–1380, 2013.

[15] R. Allez and J. Bouchaud, "Eigenvector dynamics: General theory and some applications," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 86, no. 4, Article ID 046202, 2012.

[16] M. S. Pepe and M. L. Thompson, "Combining diagnostic test results to increase accuracy," *Biostatistics*, vol. 1, no. 2, pp. 123–140, 2000.

[17] M. S. Pepe, T. Cai, and G. Longton, "Combining predictors for classification using the area under the receiver operating characteristic curve," *Biometrics: Journal of the International Biometric Society*, vol. 62, no. 1, pp. 221–229, 319, 2006.

[18] M.-J. Hsu and H.-M. Hsueh, "The linear combinations of biomarkers which maximize the partial area under the ROC curves," *Computational Statistics*, vol. 28, no. 2, pp. 647–666, 2013.

[19] W. Yu and T. Park, "Two simple algorithms on linear combination of multiple biomarkers to maximize partial area under the ROC curve," *Computational Statistics & Data Analysis*, vol. 88, pp. 15–27, 2015.

[20] C. Liu, A. Liu, and S. Halabi, "A min-max combination of biomarkers to improve diagnostic accuracy," *Statistics in Medicine*, vol. 30, no. 16, pp. 2005–2014, 2011.

[21] P. J. Heagerty and M. S. Pepe, "Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 48, no. 4, pp. 533–551, 1999.

[22] X. Huang, G. Qin, and Y. Fang, "Optimal combinations of diagnostic tests based on AUC," *Biometrics: Journal of the International Biometric Society*, vol. 67, no. 2, pp. 568–576, 2011.

[23] J. P. Egan, *Signal Detection Theory and ROC Analysis*, Academic Press, New York, NY, USA, 1975.

[24] D. D. Dorfman, K. S. Berbaum, C. E. Metz, R. V. Lenth, J. A. Hanley, and H. A. Dagga, "Proper receiver operating characteristic analysis: the bigamma model," *Academic Radiology*, vol. 4, no. 2, pp. 138–149, 1997.

[25] D. Faraggi, B. Reiser, and E. F. Schisterman, "ROC curve analysis for biomarkers based on pooled assessments," *Statistics in Medicine*, vol. 22, no. 15, pp. 2515–2527, 2003.

[26] Y. Huang and M. S. Pepe, "A parametric ROC model-based approach for evaluating the predictiveness of continuous markers in case-control studies," *Biometrics: Journal of the International Biometric Society*, vol. 65, no. 4, pp. 1133–1144, 2009.

[27] J. A. Hanley, "Receiver operating characteristic (ROC) methodology: the state of the art," *Critical Reviews in Computed Tomography*, vol. 29, no. 3, pp. 307-35, 1989.

[28] R. B. Nelsen, *An Introduction to Copulas*, vol. 139 of *Lecture Notes in Statistics*, Springer, Berlin, Germany, 1999.

[29] B. I. Rini, S. Halabi, J. E. Rosenberg et al., "Bevacizumab plus interferon alfa compared with interferon alfa monotherapy in patients with metastatic renal cell carcinoma: CALGB 90206," *Journal of Clinical Oncology*, vol. 26, no. 33, pp. 5422–5428, 2008.

[30] B. I. Rini, S. Halabi, J. E. Rosenberg et al., "Phase III trial of bevacizumab plus interferon alfa versus interferon alfa monotherapy in patients with metastatic renal cell carcinoma: final results of CALGB 90206," *Journal of Clinical Oncology*, vol. 28, no. 13, pp. 2137–2143, 2010.

[31] A. B. Nixon, S. Halabi, I. Shterev et al., "Identification of predictive biomarkers of overall survival (OS) in patients (pts) with advanced renal cell carcinoma (RCC) treated with interferon alpha (I) +/- bevacizumab (B): Results from CALGB 90206 (Alliance)," *Journal of Clinical Oncology*, vol. 31, article no. 4520, 2013.

[32] A. B. Nixon, H. Pang, M. D. Starr et al., "Prognostic and predictive blood-based biomarkers in patients with advanced pancreatic cancer: Results from CALGB80303 (alliance)," *Clinical Cancer Research*, vol. 19, no. 24, pp. 6957–6966, 2013.

[33] L. H. Cox, M. M. Johnson, and K. Kafadar, "Exposition of statistical graphics technology," in *Proceedings of the ASA Proceedings of the Statistical Computation Section*, pp. 55-56, 1982.

[34] J. B. Copas and P. Corbett, "Overestimation of the receiver operating characteristic curve for logistic regression," *Biometrika*, vol. 89, no. 2, pp. 315–331, 2002.

[35] N. Erho, A. Crisan, I. A. Vergara et al., "Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy," *PLoS ONE*, vol. 8, no. 6, Article ID e66855, 2013.

[36] J. A. Sparano, R. J. Gray, D. F. Makower et al., "Prospective validation of a 21-gene expression assay in breast cancer," *The New England Journal of Medicine*, vol. 373, no. 21, pp. 2005–2014, 2015.

Hindawi
Journal of Probability and Statistics
Volume 2019, Article ID 8057820, 9 pages
https://doi.org/10.1155/2019/8057820

Research Article

Atrial Fibrillation Detection by the Combination of Recurrence Complex Network and Convolution Neural Network

Xiaoling Wei,[1] Jimin Li,[2] Chenghao Zhang,[3] Ming Liu,[1] Peng Xiong,[1] Xin Yuan,[1] Yifei Li,[1] Feng Lin,[4] and Xiuling Liu[1]

[1] *Key Laboratory of Digital Medical Engineering of Hebei Province, College of Electronic and Information Engineering, Hebei University, Baoding, China*
[2] *College of Cyber Security and Computer, Hebei University, Baoding, China*
[3] *Department of Applied Mathematics, School of Natural and Applied Sciences, Northwestern Polytechnical University, Xi'an, China*
[4] *Nanyang Technological University, Singapore*

Correspondence should be addressed to Jimin Li; 1026724993@qq.com, Ming Liu; gleer@126.com, Peng Xiong; xiongde.youxiang@163.com, and Xiuling Liu; liuxiuling121@hotmail.com

Received 3 July 2018; Accepted 7 November 2018; Published 3 January 2019

Guest Editor: Min Zhang

In this paper, R wave peak interval independent atrial fibrillation detection algorithm is proposed based on the analysis of the synchronization feature of the electrocardiogram signal by a deep neural network. Firstly, the synchronization feature of each heartbeat of the electrocardiogram signal is constructed by a Recurrence Complex Network. Then, a convolution neural network is used to detect atrial fibrillation by analyzing the eigenvalues of the Recurrence Complex Network. Finally, a voting algorithm is developed to improve the performance of the beat-wise atrial fibrillation detection. The MIT-BIH atrial fibrillation database is used to evaluate the performance of the proposed method. Experimental results show that the sensitivity, specificity, and accuracy of the algorithm can achieve 94.28%, 94.91%, and 94.59%, respectively. Remarkably, the proposed method was more effective than the traditional algorithms to the problem of individual variation in the atrial fibrillation detection.

1. Introduction

Atrial fibrillation (AF) is the most common type of cardiac arrhythmia in clinical setting, affecting about 1–2% of the general population [1]. Clinical progress indicates that the presence of AF is associated with an increased risk for stroke, heart failure, hospitalization, and death [2]. However, the occurrence of AF is usually unknown because for many patients, the condition is asymptomatic and thus remains undetected. As a result, there is a pressing need to develop AF detection methods.

Electrocardiogram (ECG) is commonly used as a diagnostic tool for AF detection, and considerable research has been conducted on ECG. These works are either based on RR interval (RRI, i.e., the interval between two adjacent QRS complex waves) variability or abnormal atrial activity (AA) (AlGhatri [3]). Previous results showed that the RRI-based algorithms are robust compared with the AA-based algorithms (Kikillus [4] and Dash [5]). However, such methods failed to be effective if the patient has a pacemaker, is taking rate-control drugs, or has other simultaneous heart problems, such as atrioventricular (AV) block [6]. Thus, it is necessary to develop AF detection algorithms based on the AA feature, namely, designing rate-independent methods [7].

In view of the atrial activity, during AF, the P-wave is replaced by fibrillatory waves. Thus, a natural way to detect AF is to check the absence of P-waves. Previous algorithms were proposed to address this issue [8–10]; however, the results were not satisfactory because P-wave fiducial point detection is challenging, especially for dynamic monitoring applications.

Recently, signal processing techniques have been employed to extract AA features from ECG waves for AF detection. Stridh *et al.* proposed using a time-frequency distribution estimation method to estimate the fibrillation frequency of the ECG signal, in which a set of parameters describing the fundamental frequency, amplitude, shape, and signal-to-noise ratio of the atrial waveforms are derived based on the frequency-shift of an adaptively updated spectral profile [11]. Lee *et al.* analyzed the dominant frequency of the atrial activity by using the variable frequency complex demodulation (VFCDM) method [12]. The value of the dominant frequency has been shown to be a distinctive feature for AF detection. In another ECG-based pattern analysis method for the classification of normal sinus rhythm and atrial fibrillation (AF) beats [13], the denoised and registered ECG beats were subjected to independent component analysis (ICA) for data reduction, and the ICA weights were used as features for classification using Naive Bayes and Gaussian mixture model (GMM) classifiers. All of these methods use handcraft features for pattern recognition. Such features are not invariant on different personalities. In their experiments, the classification accuracy was estimated by tenfold cross-validation, where the probability that the training set contains training samples of every user who provides test ECG samples is great. We use the term individual variation to refer to the above phenomena. However, in ECG monitoring applications, it is crucial that the system is able to tackle this problem.

Magnitude-squared coherence, a frequency domain measure of the linear phase relation between two signals, has been shown to be a reliable discriminator of AF [14, 15]. However, the accuracy of the corresponding AF detection algorithm is relatively low; thus, it has to be combined with the RRI feature in order to achieve acceptable accuracy. The Recurrence Complex Network has been employed to detect AF from dog epicardial signals recorded by an epicardial mapping system with 128 unipolar electrodes [16]. It has been demonstrated that the phase space of the Recurrence Complex Network is suitable for between distinguishing normal sinus rhythm and atrial fibrillation beats. However, in [16], only two numerical features calculated from the adjacent matrix of the complex network are used to detect AF. This process may cause the loss of a lot of discriminating information of the adjacent matrix.

The objective of this paper is to improve the performance of the AF detection algorithm by combining the Recurrence Complex Network (RCN) with convolution neural network (CNN). As one of the deep learning algorithms [17], CNN has great potential in feature extraction and has been applied to image processing and speech recognition with notable success [18–20]. In the proposed algorithm, CNN is exploited to learn robust AF features from the output of the RCN and then to detect AF signal with high accuracy. The proposed AF detection algorithm is composed of two procedures. The first is a heartbeat classification procedure that can distinguish between AF beat and normal beat based on the ECG waveform of a single heartbeat. The second is a voting procedure that improves classification performance by fusing the classification results of multiple beats. The first procedure is the crucial part, in which the synchronization feature of each heartbeat is first extracted by the RCN, and then, a CNN is used to extract more abstract AF features and recognize an AF heartbeat. Experimental results on the MIT-BIH database show that the AF features learned by the CNN are robust to the variation of the ECG signals between different personalities so that the proposed algorithm has good generalization ability.

2. The Data

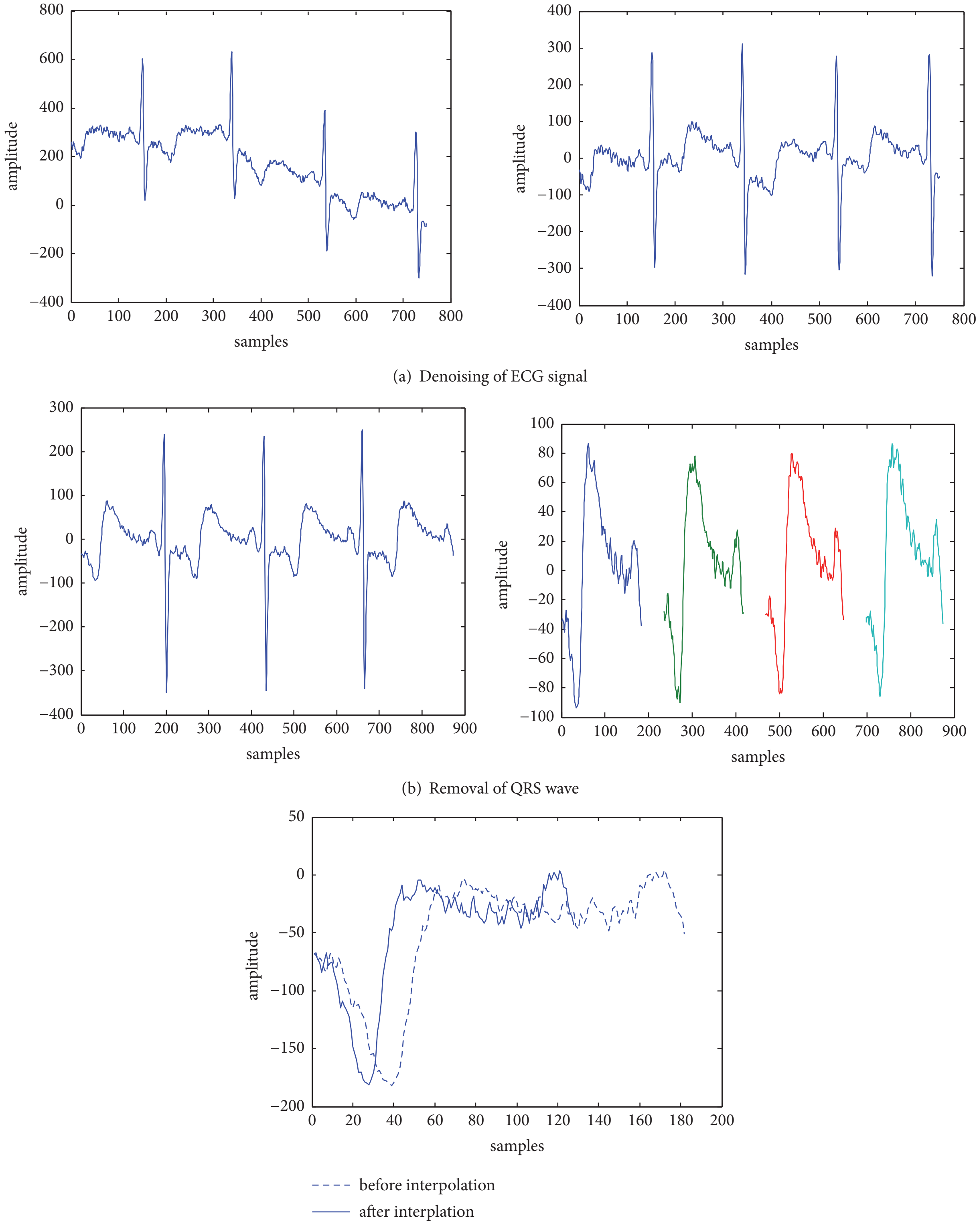
The real data (surface ECGs) used in this method were provided by the MIT-BIH AF database (AFDB) [21]. The database is from Physionet [22] and includes 25 long-term (10 hour) annotated ECG recordings of humans with AF and contains 299 AF episodes. Each recording contains two ECG signals (ECG1 and ECG2), which are sampled at 250HZ and 12-bit resolution. In this work, only ECG1 signals are used to evaluate the AF detection methods.

3. Methods

3.1. Data Preprocessing. For each data recording, a seven-order Butterworth bandpass filter is applied with poles at 0.5 Hz and 49 Hz to reduce baseline wander (BW) and noise. Then, the onset of the QRS wave is detected by finding the local maximums of the convolution between the ECG recording and a set of predefined QRS models. At each QRS onset point, the QRS wave is canceled based on the most matched model. The remaining signals are departed into segments; each of which is approximately the AA segment of a heartbeat. All the segments are interpolated into 128 bit data samples with the Fourier transform interpolation. Next, an AF detection algorithm is developed based on such samples. The main ECG preprocessing steps are illustrated in Figure 1, which clearly illustrates the changing process of the data. Figure 1(a) shows that the Butterworth filter can successfully correct the baseline and reduce the effects of the noise. In Figure 1(b), the right figure only contains information outside the QRS wave. It shows that the ventricular signals are almost removed; thus, the output signal essentially represents the AA signal. These are the single heartbeats before and after the interpolation operation, as shown in Figure 1(c).

3.2. Extracting Low Level AF Features Based on the Recurrence Complex Network. The ECG data is a nonstationary time series [23]; thus, it can be analyzed by the Recurrence Complex Network (RCN), a popular tool for processing nonstationary time series [23, 24]. Traditionally, there are two issues that need to be explored when applying the RCN: the construction of the recurrence matrix and the extraction of the RCN features. This section mainly focuses on the construction of the recurrence matrix from the ECG data.

The recurrence matrix is obtained by the phase space reconstruction method. Generally, there are two kinds of phase space construction methods: the time delay method and the derivative reconstruction method [24]. In this study, the time delay method was selected because the derivation is sensitive to the calculation error. Let $x(t) = \{x(t_1)\ \ x(t_2)\ \ \cdots\ \ x(t_N)\}$ denote an ECG data of length N, and then, the vector $X(t_i) = [x(t_i)\ \ x(t_{i+\tau})\ \ \cdots\ \ x(t_{i+(m-1)\tau})]$ represents a vector in the phase space. Here, m is the



(a) Denoising of ECG signal

(b) Removal of QRS wave

(c) Interpolation of the AA segment

Figure 1: ECG data preprocessing.

embedding dimension, and τ is the embedding delay time. If the parameters m and τ are properly specified, the dynamic characteristics of the data will be transferred into the relationship between the vectors in the phase space, and it will be much easier to observe and extract the dynamic features of the data than that in the original space.

The most common method for choosing the time delay parameter τ is based on the mutual information between the coordinates of the phase space (Frase [25]). By the assignment $[s_i, q_i] = [x(t_i), x(t_{i+\tau})]$, a couple of random variables S and Q are defined, where $s_i = x(t_i)$ is an instance of S, and $q_i = x(t_{i+\tau})$ is an instance of Q. The average amount of information gained from a specific value of S, named the entropy $H(S)$, is defined as the following:

$$H(S) = -\sum_i P_S(s_i) \log P_S(s_i), \tag{1}$$

where $P_S(s_i)$ is the probability that the observed value of the random variable S is s_i.

The entropy $H(Q)$ is defined in the same way. Moreover, the joint entropy of the couple $[S, Q]$ can be defined as

$$H(S,Q) = -\sum_{i,j} P_{S,Q}\left(s_i, q_j\right) \log P_{S,Q}\left(s_i, q_j\right), \tag{2}$$

where $P_{S,Q}(s_i, q_j)$ is the probability that the observed values of S are s_i, and Q is q_j.

Then, the mutual information between S and Q can be defined as $I(S,Q)$:

$$I(S,Q) = H(S) + H(Q) - H(S,Q), \tag{3}$$

In recalling the definitions of the random variables S and Q, it can be determined that $I(S,Q)$ is a function of τ. The research work of [25] demonstrated that the proper value of the time delay τ corresponds with the first local minimum of $I(S,Q)$.

The problem of determining the embedding dimension m was explored in depth in [26], in which an efficient method for determining the embedding dimension m was developed based on the fact that a low embedding dimension results in points that are far apart in the high dimensional phase space being moved closer together in the reconstructed space [27]. This method was adopted in our AF detection algorithm, and in the following part, we briefly review it.

Consider again the ECG data $x(t) = \{x(t_1)\ x(t_2)\ \cdots\ x(t_N)\}$. Suppose that the vectors constructed with dimension d are $X_d(t_i) = [x(t_i)\ x(t_{i+\tau})\ \cdots\ x(t_{i+(d-1)\tau})]$, where $i = 1, 2, \cdots N-(d-1)\tau$, and the nearest neighbor of $X_d(t_i)$ in the phase space is $X_d(t_{n(i,d)})$. The vectors constructed with dimension $d+1$ are $X_{d+1}(t_i) = [x(t_i)\ x(t_{i+\tau})\ \cdots\ x(t_{i+d\tau})]$, where $i = 1, 2, \cdots N - d\tau$. Then, a quantity measure of the difference of the two phase spaces in view of the distances between the adjacent vectors is defined as

$$a(i,d) = \frac{\left\|X_d(t_i) - X_d\left(t_{n(i,d)}\right)\right\|}{\left\|X_{d+1}(t_i) - X_{d+1}\left(t_{n(i,d)}\right)\right\|}, \quad i = 1, 2, \cdots N - d\tau \tag{4}$$

where $\|\cdot\|$ is a measurement of the Euclidian distance, such as the maximum norm. According to Cao, etc. [26], the parameter d can be determined by the function:

$$E1(d) = \frac{(N - d\tau)\sum_{i=1}^{N-(d+1)\tau} a(i, d+1)}{(N - (d+1)\tau)\sum_{i=1}^{N-d\tau} a(i,d)} \tag{5}$$

Having chosen appropriate parameters for the phase space, the dynamic character of the original data can be represented by the following $L \times L$ recurrence matrix R:

$$R(i,j) = \left\|X(t_i) - X\left(t_j\right)\right\|, \quad i, j = 1, 2, \cdots L \tag{6}$$

where $L = N - (m-1)\tau$. The matrix R is a symmetric matrix with the diagonal element of 0. In our algorithm, N=128, $\tau = 4$, and $m = 10$ (see Section 4.1), such that $L = 92$.

Traditionally, the recurrence matrix is binarilized, and some numerical features are extracted through the manual method. Then, the input samples can be classified with algorithms such as fuzzy c-means (FCM). However, it is difficult to manually define the appropriate features for the ECG data. To solve this problem, we propose to extract features from the recurrence matrix automatically by using the convolution neural network (CNN). Firstly, we calculate the eigenvalues of the recurrence matrix, and then, they are sent into the CNN. The CNN extracts the features and classifies the data. The eigenvalues of each data sample form a 92-byte feature vector.

3.3. AF Detection Based on the Convolutional Neural Network

3.3.1. Architecture of the CNN. The convolutional neural network (CNN) addresses the feature learning problem through the calculation of multiple levels of data representations by the operation involved in the multiple layers of the CNN. Except for the first layer and the top layer, the main part of the CNN is composed of alternating layers of convolution and pooling.

As illustrated in Figure 2, the convolution layer adopted in this study consists of a group of fully connected feature maps C_j $(j = 1, 2, \cdots, N_o)$ (assume N_o is the total number). Each feature map C_j is obtained by a summation of the convolutions from all the input feature maps (assume N_I is the total number), O_i $(i = 1, 2, \cdots, N_I)$, and a series of weight vectors $W_{i,j}$, $i = 1, 2, \cdots, N_I$, $j = 1, 2, \cdots, N_O$, i.e.,

$$C_j = \sigma\left(\sum_{i=1}^{N_I} O_i * \mathbf{W}_{i,j}\right), \quad j = 1, 2, \cdots, N_o \tag{7}$$

where $\sigma(x)$ is a nonlinear activation function, for example, $\sigma(x) = 1/(1 + \exp(-x))$. The term feature map is borrowed from image processing applications, in which the input and output of each layer of the CNN are 2-dimensional arrays. However, the input of the CNN is a 1-dimensional vector, and as a result, all of the feature maps are 1-dimensional vectors in our AF detection algorithm.

The weight vectors of the convolutional layer, $W_{i,j}$, $i = 1, 2, \cdots, N_i$, $j = 1, 2, \cdots, N_O$, can be seen as trainable feature

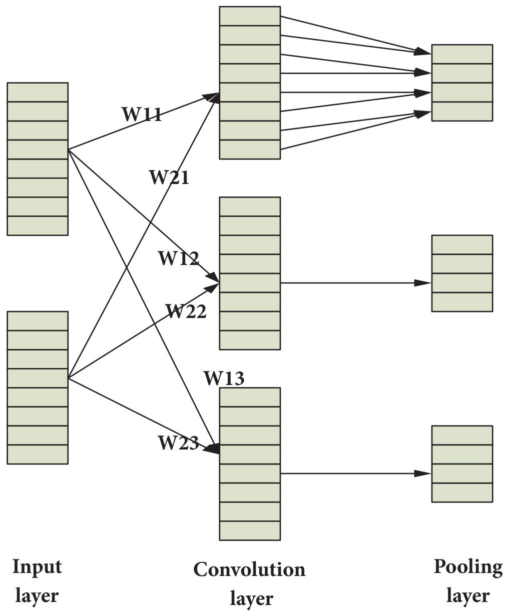


Figure 2: Illustration of a convolution layer and the subsequent pooling layer.

extraction operators; each of which enhances one kind of feature and weakens the others. When the CNN is trained with sufficient training data, the feature maps, which are obtained by using these weight vectors, will be turned into an appropriate representation for recognition of the input data.

Each convolutional layer is followed by a pooling layer, as shown in Figure 2. The pooling layer is also composed of feature maps. Each feature map P_j $(j = 1, 2, \cdots, N_o)$ in the pooling layer is obtained by applying a pooling operation to the units of a convolution layer feature maps C_j $(j = 1, 2, \cdots, N_o)$. There are usually two kinds of pooling operations: maximization pooling and averaging pooling. Here, averaging pooling was adopted, which is defined as

$$p_{j,m} = r \sum_{n=1}^{G} c_{j,(m-1)\times s+n} \tag{8}$$

where $p_{j,m}$ is the m-th unit of P_j; $c_{j,k}$ is the k-th unit of C_j; G is the pooling size, which determines a pooling window; s is the shift size, which determines the overlap of adjacent pooling windows; and r is the scaling factor, which is selected as one in G. By the pooling operation, the resolutions of the feature maps are reduced so that the features learned by the CNN are robust to small variations in location.

The CNN used for AF detection has six layers (as illustrated in Figure 3). It consists of one input layer; two convolutional layers, which are denoted as C1 and C2, respectively; two pooling layers which are denoted as S1 and S2, respectively; and one output layer. The 92 eigenvalues of the reconstructed recurrence matrix form the ECG sample of a heartbeat and are mounted into the input layer. The output layer has only two units o_n, where $n = 1, 2$; o_1 corresponds with the normal heat beat class, and o_2 corresponds with the AF heart beat class. Suppose that the output units are denoted as the units in the final pooling layer (assume N_F is the total number) by $p_{F,m}$, where $m = 1, 2, \cdots, N_F$, and the weight between $p_{F,m}$ and o_n is $w_{m,n}$, where $m = 1, 2, \cdots, N_F$, $n = 1, 2$. Then, the final outputs can be calculated as follows:

$$o_n = \frac{\exp\left(\sum_{m=1}^{N_F} p_{F,m} w_{m,n}\right)}{\sum_{n'=1}^{2} \exp\left(\sum_{m=1}^{N_F} p_{F,m} w_{m,n'}\right)} \tag{9}$$

The CNN can be trained with the back-propagation (BP) algorithm with the loss function:

$$E_{x_i}(\theta) = -\sum_{n=1}^{2} y_{i,n} o_n \tag{10}$$

where θ denotes all of the weights of the CNN, x_i is an input sample, and $y_i = \begin{bmatrix} y_{i,1} & y_{i,2} \end{bmatrix}$ is the binary encoding vector target label for x_i. Details of the training algorithm are available in [28].

The number of feature maps in each convolutional layer and the pooling parameters are chosen experimentally in Section 4.

3.3.2. AF Detection. The input ECG data is preprocessed and segmented into 128-bit samples, where each sample corresponds to the atrial activity (AA) signal of one heartbeat. Then, the recurrence matrix is calculated. The eigenvalues of the recurrence matrix, which form a 92-byte feature vector, are sent to a CNN. The details of the CNN are introduced in the following.

C1 layer: The C1 layer is a convolution layer. It consists of six feature maps with a vector of 1*80. Each unit of one feature map in this layer obtains the input from a local area. The size of the convolution kernel determines the size of the receptive field of neurons. Therefore, it is important to set an appropriate convolution kernel size. Here, the convolution kernel is set to be 13, and the size of the output feature map is 80(92-13+1=80). The inner information of the input data is extracted through different convolution kernels.

S2 layer: The S2 layer is a pooling layer. The obtained feature from C1 is sampled according to the principle of local image characteristics. The sampling is achieved by using a pooling function to several units in a region of a size determined by the pooling size parameter. After the experiment, the size is set as 2. Therefore, the size of the obtained feature map in this layer is 40 (80/2=40). The further feature extraction will cause it to be invariant to small variations in location. The resolution of the obtained feature map is reduced, but most of the information is retained.

C3 layer: The C3 layer is similar to that of C1. The size of the obtained feature map is 28 (40-13+1=28). As mentioned above, the pooling layer increases the receptive field of neurons. Therefore, a better feature structure is acquired for the depth structure.

S4 layer: This layer is the same as the S2 layer. The size of the feature maps is 14 (28/2=14).

Output layer: The output layer is fully connected to S4 layer. The number of S4 neurons is 12*14=168. Each neuron is connected to the output. There are 168*2=336 connections

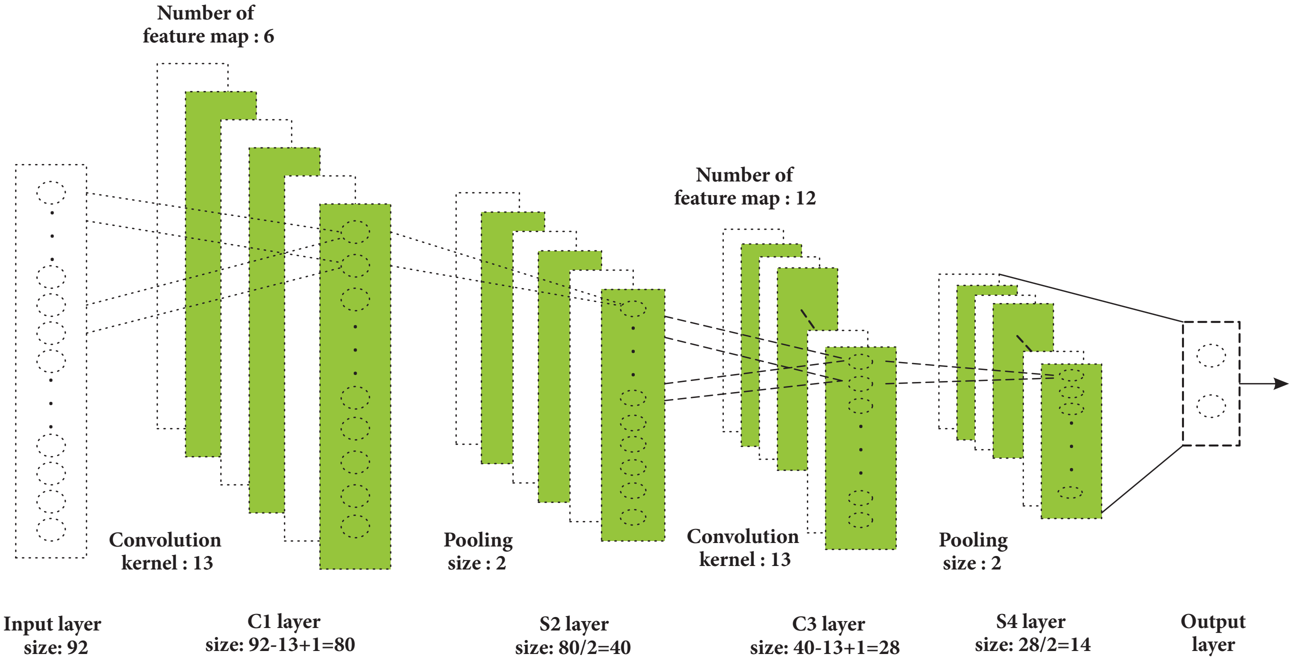


FIGURE 3: Structure diagram of the CNN.

because the output layer consists of two neurons. The output will be closer to the desired output after several times of training through the BP algorithm to update the weights of the network.

3.4. Majority Voting. Although the beat-wise AF detection algorithm is important in exploring the underlying feature of AF, its classification accuracy is relatively low. To improve algorithm performance, the majority voting methodology was adopted. Before AF detection, the ECG data is segmented into beat-wise data samples. Each adjacent G sample is used as a collective candidate for AF detection. The samples of one candidate are classified using the above method, and then, the classification results are integrated by majority voting to determine whether it is AF data. The parameter G will be determined experimentally in the next section.

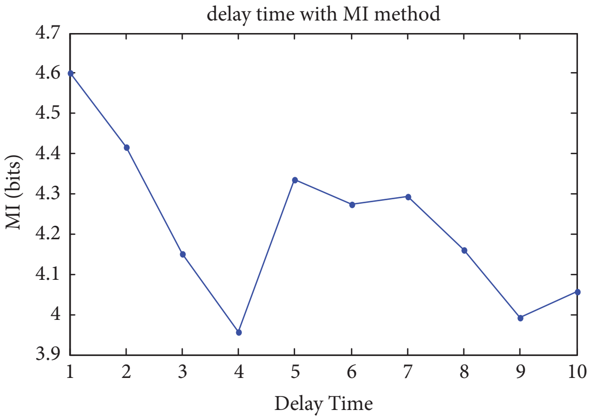


FIGURE 4: Selection of the delay time.

4. Experiments and Discussion

All programs and graphs were created in Matlab (R2015b version 8.6.0.267246, Mathworks). The 23 recordings in the database were divided into two groups. The first group contains 15 recordings, and the second group contains 8 recordings. The recordings of the two groups were obtained from different subjects. From the first group, 120,000 NSR (Normal sinus rhythm) heartbeat AA data samples and 120,000 AF samples, respectively, were obtained with the preprocessing method detailed in Section 3.1. All of the 240,000 samples were used to construct the training set. From the second group, 40,000 NSR heartbeat AA data samples and 40,000 AF samples, respectively, were obtained with the same preprocessing method. These 80,000 samples were used to construct the testing set. The goal of such an arrangement is to test whether the AF detection algorithms can be adapted to different individuals.

4.1. Choice of Time Delay and Embedding Dimension. There are two parameters of the reconstructed phase space that need to be determined: the delay time and the embedding dimension. Figure 4 plots the MI versus the delay time τ. The delay time corresponding to the first local minimum of MI ($\tau = 4$) is selected for the phase space. Figure 5 plots the function $E1(d)$. It can be seen that when the dimension exceeds 9, the $E1$ value is close to 1 and does not significantly change with an increased embedding dimension. Based on Cao's method, m is set as 10.

Table 1: Classification rates of CNN under different lengths of the convolution kernel.

length of convolution kernel	5	9	13	17
training set	75.68%	77.17%	82.37%	78.3%
testing set	74.39%	73.5%	80.77%	71.2%

Table 2: Classification rates of CNN under different number of feature maps.

N_{c3}	N_{c1}							
	3		6		9		12	
	train	test	train	test	train	test	train	test
3	74.28%	77.29%	76.95%	75.29%	80.13%	74.94%	76.87%	79.75%
6	73.23%	76.96%	82.19%	77.42%	82.45%	74.95%	73.92%	59.95%
9	71.71%	74.74%	78.08%	79.89%	80.91%	75.65%	82.59%	79.65%
12	74.66%	71.1%	82.37%	80.77%	79.94%	71.11%	82.2%	76.31%

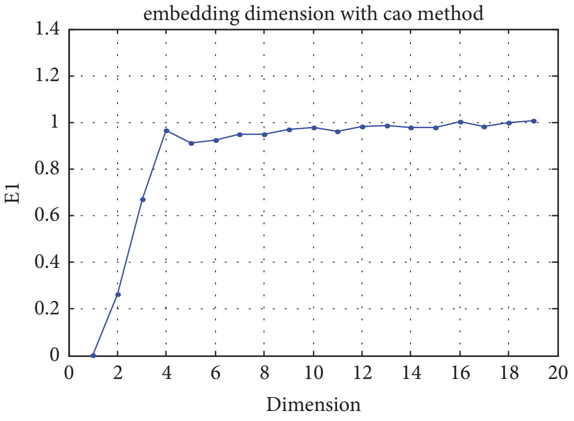


Figure 5: Selection of the embedding dimension.

4.2. The Effects of Varying CNN Parameters. In order to select the best parameters for the CNN, the performance of the CNN is evaluated using different parameters.

(1) Effects of Different Convolution Kernel Lengths. There are four parameters that need to be determined: the pooling size, the length of the convolution kernels, the number of feature maps in the C1 layer, and the number of feature maps in the C3 layer. In the present algorithm, the length of the input vector of the CNN is not too large; thus, the big pooling size may result in information loss. Therefore, the pooling size is fixed at 2.

As an initiation, the number of feature maps in the C1 layer and C2 layer is set as 6 and 12, respectively, according to ref. [18], and the effects of different lengths of convolution kernels are observed. Table 1 shows the classification rates of the CNN under different convolution kernel lengths. A length of 13 produced the maximal classification rate.

(2) Effects of Various Number of Feature Maps. Table 2 illustrates the accuracy of different feature maps in C1 layer (N_{c1}) and C3 layer (N_{c3}). The results reveal that the CNN performs best when $N_{c1} = 6$ and $N_{c3} = 12$.

4.3. Experimental Results of Beat-Wise AF Detection. To illustrate the effectiveness of the CNN, the CNN is compared with other popular classification methods. Three measurements are used to evaluate the methods: accuracy (AC), sensitivity (SE), and specificity (SP). The inputs of all three classifiers are the low level features obtained by the method detailed in Section 3.2. Table 3 demonstrates that the CNN greatly outperforms the others.

Most of the rate-independent AF detection algorithms are unable to solve the problem of individual variation. According to our investigation, only the Magnitude-squared coherence (MSC) algorithm [15] and the Recurrence Complex Network (RCN) algorithm [16] can recognize the samples of different individuals based on beat-wise AA samples. Table 4 presents a comparison of the proposed beat-wised AF detection algorithm (BWAD) with these two algorithms. For the contrast experiments, all of the ECG recordings are preprocessed with the method described in Section 3.1, and, the training and testing sets are constructed as previously described. In the BWAD algorithm, the low level features are first extracted and then, the CNN is used to extract the high-level features and classify them. As for the MSC algorithm, the feature vectors are calculated between each data sample and the previous sample, and the samples are classified based on a hand measurement that is detailed in [15]. For the RCN algorithm, the recurrence matrix is calculated the same as that in the proposed algorithm, and, the samples are classified based on two hand measurements that are detailed in [16].

It can be seen that the proposed BWAD algorithm outperforms the traditional algorithms. Traditional algorithms perform poorly in beat-wise rate-independent AF detection because they rely on manually obtained features. In contrast, the BWAD algorithm effectively solves this problem by using CNN to extract high-level features for classification.

4.4. Experimental Results of Majority Voting. The performance of the proposed algorithm can be improved by majority voting, in which the outputs of G adjacent heartbeat samples are integrated to obtain an accurate result. Table 5

Table 3: Comparison of CNN with typical classification methods.

method	AC	SE	SP
SVM	61.34%	56.25%	66.37%
KNN	70.95%	88.97%	52.93%
CNN	80.77%	89.18%	72.37%

Table 4: Comparison with traditional rate-independent AF detection algorithms.

methods	AC	SE	SP
MSC	66.31%	71.54%	61.09%
RCN	60.03%	66.16%	53.91%
BWAD	80.77%	89.18%	72.37%

Table 5: Results of majority voting under different parameters.

G	AC	SE	SP
13	92.92%	94.81%	90.99%
15	92.97%	96.06%	89.87%
17	94.09%	94.30%	93.88%
19	94.12%	95.72%	92.52%
21	94.59%	94.28%	94.91%

Table 6: Comparison of the time spent in each process.

Remove QRS wave and noise reduction	The interpolation process	RCN extracting feature process	Testing process
0.11 seconds	0.00023 seconds	0.0075 seconds	0.00092 seconds

lists the classification rates of the voting algorithm under different parameters (G). It can be seen that a larger G value usually leads to a better performance.

4.5. The Calculation of the Complexity. The configuration of the computer used for the program is an Intel Pentium Dual-Core with a processor speed of 2.2GHz and a memory size of 3.18GB. For the proposed algorithm, training the CNN is a time-consuming process. However, the training process can be carried out off-line. The training process (i.e., the whole data preprocessing process and the CNN training process (10 times)) requires approximately 9.65 hours for the 24,000 samples. Table 6 lists the results.

As for the testing process, it was determined that the process of removing the QRS wave and reducing the noise was the most time-consuming process. After several experiments, the time spent in each process for one sample was obtained, and it was revealed that the testing process is about 0.1186 seconds for a sample. Therefore, this method can be used in real-time signal processing.

5. Conclusion

In this paper, a novel rate-independent AF detection algorithm that combines RCN and CNN based on AA features is presented. Firstly, the recurrence matrix is calculated with RCN, and the eigenvalues of the matrix are extracted to characterize atrial activity. Then, CNN is employed, which leverages the multilayer structures and presents an increasingly abstract representation of the input. These signals are distinguished through the optimization of the network so as to extract high-level features and classify the input sample. Finally, majority voting is utilized to improve algorithm performance.

In the experiments, the training set and testing set are constructed with a special arrangement so that the data samples of each set are obtained from different subjects. The proposed algorithm achieves an accuracy of 94.59%, which is comparable to popular RRI-based methods. Moreover, the proposed rate-independent algorithm is applicable to patients with rate-controlled drugs or pacemakers. Furthermore, the developed method solves the problem of individual variation. Therefore, it is evident that the proposed method can detect AF with superior performance.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by National Natural Science Foundation of China (61473112), Foundation for Distinguished Young Scholars of Hebei Province (F2016201186), Natural Science Foundation of Hebei Province (F2015201112), and Science and Technology Research Project for Universities and Colleges in Hebei Province (ZD2015067).

References

[1] C. Bruser, J. Diesel, H. Zink et al., "Automatic detection of atrial fibrillation in cardiac vibration signals," *IEEE Journal of Biomedical & Health Informatics*, vol. 17, no. 1, pp. 162–171, 2013.

[2] G. Pagana, L. Galleani, S. Gross et al., "Time-frequency analysis of the endocavitarian signal in paroxysmal atrial fibrillation," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 10, article no. 20, pp. 2838–2844, 2012.

[3] M. AlGhatrif and J. Lindsay, "A brief review: history to understand fundamentals of electrocardiography," *Journal of Community Hospital Internal Medicine Perspectives (JCHIMP)*, vol. 2, no. 1, p. 14383, 2012.

[4] N. Kikillus, G. Hammer, N. Lentz, F. Stockwald, and A. Bolz, "Three different algorithms for identifying patients suffering from atrial fibrillation during atrial fibrillation free phases of the ECG," in *Proceedings of the Computers in Cardiology 2007, CAR 2007*, pp. 801–804, USA, October 2007.

[5] S. Dash, K. H. Chon, S. Lu, and E. A. Raeder, "Automatic real time detection of atrial fibrillation," *Annals of Biomedical Engineering*, vol. 37, no. 9, pp. 1701–1709, 2009.

[6] S. Ladavich and B. Ghoraani, "Developing an atrial activity-based algorithm for detection of atrial fibrillation," in *Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*, pp. 54–57, USA, August 2014.

[7] S. Ladavich and B. Ghoraani, "Rate-independent detection of atrial fibrillation by statistical modeling of atrial activity," *Biomedical Signal Processing and Control*, vol. 18, pp. 274–281, 2015.

[8] A. Petrenas, L. Sornmo, V. Marozas, and A. Lukosevicius, "A noise-adaptive method for detection of brief episodes of paroxysmal atrial fibrillation," in *Proceedings of the 2013 40th Computing in Cardiology Conference, CinC 2013*, pp. 739–742, Spain, September 2013.

[9] S. Babaeizadeh, R. E. Gregg, E. D. Helfenbein, J. M. Lindauer, and S. H. Zhou, "Improvements in atrial fibrillation detection for real-time monitoring," *Journal of Electrocardiology*, vol. 42, no. 6, pp. 522–526, 2009.

[10] R. Couceiro, P. Carvalho, J. Henriques, M. Antunes, M. Harris, and J. Habetha, "Detection of Atrial Fibrillation using model-based ECG analysis," in *Proceedings of the 2008 19th International Conference on Pattern Recognition (ICPR)*, pp. 1–5, Tampa, FL, USA, December 2008.

[11] M. Stridh, L. Sörnmo, C. J. Meurling, and S. B. Olsson, "Sequential Characterization of Atrial Tachyarrhythmias Based on ECG Time-Frequency Analysis," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 1, pp. 100–114, 2004.

[12] J. Lee, D. D. McManus, P. Bourrell, L. Sörnmo, and K. H. Chon, "Atrial flutter and atrial tachycardia detection using Bayesian approach with high resolution time-frequency spectrum from ECG recordings," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 992–999, 2013.

[13] R. J. Martis, U. R. Acharya, H. Prasad, C. K. Chua, and C. M. Lim, "Automated detection of atrial fibrillation using Bayesian paradigm," *Knowledge-Based Systems*, vol. 54, pp. 269–275, 2013.

[14] L. S. Sarraf, J. A. Roth, and K. M. Ropella, "Differentiation of atrial rhythms from the electrocardiogram with coherence spectra," *Journal of Electrocardiology*, vol. 35, no. 1, pp. 59–67, 2002.

[15] J. Lee, Y. Nam, D. D. McManus, and K. H. Chon, "Time-varying coherence function for atrial fibrillation detection," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2783–2793, 2013.

[16] Y. Zhang, Y. Wang, C. Yang, X. Wu, and Y. Qin, "Atrial fibrillation detection using spectra of FSD recurrence complex network," in *Proceedings of the 4th International Conference on Audio, Language and Image Processing, ICALIP 2014*, pp. 44–47, China, July 2014.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR '12)*, pp. 3288–3291, IEEE, November 2012.

[19] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 4353–4361, USA, June 2015.

[20] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[21] G. B. Moody and R. R. Mark, "New method for detecting atrial fibrillation using R-R intervals," in *Proceedings of the Computers in Cardiology, 10th Annual Meeting*, pp. 227–230, 1983.

[22] A. L. Goldberger, L. A. Amaral, L. Glass et al., "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals.," *Circulation*, vol. 101, no. 23, pp. E215–E220, 2000.

[23] S. Pola, A. Macerata, M. Emdin, and C. Marchesi, "Estimation of the power spectral density in nonstationary cardiovascular time series: Assessing the role of the time-frequency representations (TFR)," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 1, pp. 46–59, 1996.

[24] Z. Gao and N. Jin, "Erratum: "Complex network from time series based on phase space reconstruction" [Chaos 19, 033137 (2009)]," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 20, no. 1, p. 019902, 2010.

[25] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical Review A: Atomic, Molecular and Optical Physics*, vol. 33, no. 2, pp. 1134–1140, 1986.

[26] L. Cao, "Practical method for determining the minimum embedding dimension of a scalar time series," *Physica D: Nonlinear Phenomena*, vol. 110, no. 1-2, pp. 43–50, 1997.

[27] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Physical Review A: Atomic, Molecular and Optical Physics*, vol. 45, no. 6, pp. 3403–3411, 1992.

[28] L. C. Agba and J. H. Tucker, "Enhanced back-propagation training algorithm," in *Proceedings of the 1995 IEEE International Conference on Neural Networks. Part 1 (of 6)*, pp. 2816–2820, December 1995.

Hindawi
Journal of Probability and Statistics
Volume 2018, Article ID 2834183, 9 pages
https://doi.org/10.1155/2018/2834183

Research Article

A Note on the Adaptive LASSO for Zero-Inflated Poisson Regression

Prithish Banerjee,[1] Broti Garai,[2] Himel Mallick,[3,4] Shrabanti Chowdhury,[5] and Saptarshi Chatterjee[6]

[1] JP Morgan Chase & Co., USA
[2] NBCUniversal, USA
[3] Department of Biostatistics, Harvard T.H. Chan School of Public Health, USA
[4] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, USA
[5] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, USA
[6] Eli Lilly and Company, USA

Correspondence should be addressed to Himel Mallick; hmallick@hsph.harvard.edu

Prithish Banerjee, Broti Garai, and Himel Mallick contributed equally to this work.

Received 23 July 2018; Accepted 21 November 2018; Published 30 December 2018

Guest Editor: Ash Abebe

We consider the problem of modelling count data with excess zeros using Zero-Inflated Poisson (ZIP) regression. Recently, various regularization methods have been developed for variable selection in ZIP models. Among these, EM LASSO is a popular method for simultaneous variable selection and parameter estimation. However, EM LASSO suffers from estimation inefficiency and selection inconsistency. To remedy these problems, we propose a set of EM adaptive LASSO methods using a variety of data-adaptive weights. We show theoretically that the new methods are able to identify the true model consistently, and the resulting estimators can be as efficient as oracle. The methods are further evaluated through extensive synthetic experiments and applied to a German health care demand dataset.

1. Introduction

Modern research studies routinely collect information on a broad array of outcomes including count measurements with excess amount of zeros. Modeling such zero-inflated count outcomes is challenging for several reasons. First, traditional count models such as Poisson and Negative Binomial are suboptimal in accounting for excess variability due to zero-inflation [1, 2]. Second, alternative zero-inflated models such as the **Z**ero-**I**nflated **P**oisson (ZIP) [2] and **Z**ero-**I**nflated **N**egative **B**inomial (ZINB) [1] models are computationally prohibitive in the presence of high-dimensional and collinear variables.

Regularization methods have been proposed as a powerful framework to mitigate these problems, which tend to exhibit significant advantages over traditional methods [3, 4]. Essentially all these methods enforce sparsity through a suitable penalty function and identify predictive features by means of a computationally efficient Expectation Maximization (EM) algorithm. Among these, EM LASSO is particularly attractive due to its capability to perform simultaneous model selection and stable effect estimation. However, recent research suggests that EM LASSO may not be fully efficient and its model selection result could be inconsistent [5, 6]. This led to a simple modification of the LASSO penalty, namely, the EM adaptive LASSO (EM AL). EM AL achieves "oracle selection consistency" by allowing different amounts of shrinkage for different regression coefficients.

Previous studies have not, however, investigated the EM AL at sufficient depth to evaluate its properties under diversified and realistic scenarios. It is not yet clear, for example, how reliable the resulting parameter estimates are in the presence of multicollinearity. In particular, the actual variable selection performance of EM AL depends on the proper construction of the data-adaptive weight vector. When the features to be associated possess an inherent collinearity, EM AL is expected to produce suboptimal results, a phenomenon that is especially evident when the sample size is limited [7]. Several remedies have been suggested for linear and generalized linear models (GLMs) such as the standard error-adjusted adaptive LASSO (SEAL) [7, 8]. However, there is a lack of similar published methods for zero-inflated count regression models. In addition, complete software packages of these methods have not been made available to the community.

We address these issues by providing a set of flexible variable selection approaches to efficiently identify correlated features associated with zero-inflated count outcomes in a ZIP regression framework. We have implemented this method as AMAZonn (**A M**ulticollinearity-adjusted **A**daptive LASSO for **Z**ero-inflated C**o**u**n**t Regressio**n**). AMAZonn considers two data-adaptive weights: (i) the inverse of the maximum likelihood (ML) estimates (EM AL) and (ii) inverse of the ML estimates divided by their standard errors (EM SEAL). We show theoretically that AMAZonn is able to identify the true model consistently, and the resulting estimator is as efficient as oracle. Numerical studies confirmed our theoretical findings. The rest of the article is organized as follows. The AMAZonn method is proposed in the next section, and its theoretical properties are established in Section 3. Simulation results are reported in Section 4 and one real dataset is analyzed in Section 5. Then, the article concludes with a short discussion in Section 6. All technical details are presented in the Appendix.

2. Methods

2.1. Zero-Inflated Poisson (ZIP) Model. Zero-inflated count models assume that the observations originate either from a "susceptible" population that generates zero and positive counts according to a count distribution or from a "nonsusceptible" population, which produces additional zeros [1, 2]. Thus, while a subject with a positive count is considered to belong to the "susceptible" population, individuals with zero counts may belong to either of the two latent populations. We denote the observed values of the response variable as $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$. Following Lambert [2], a ZIP mixture distribution can be written as

$$P(y_i = k) = \begin{cases} p_i + (1 - p_i)\, e^{-\lambda_i} & \text{if } k = 0, \\ (1 - p_i)\, \dfrac{e^{-\lambda_i}\lambda_i^k}{k!} & \text{if } k = 1, 2, \ldots, \end{cases} \tag{1}$$

where p_i is the probability of belonging to the nonsusceptible population and λ_i is the Poisson mean corresponding to the susceptible population for the i^{th} individual ($i = 1, \ldots, n$). It can be seen from (1) that ZIP reduces to the standard Poisson model when $p_i = 0$. Also, $P(y_i = 0) > e^{-\lambda_i}$, indicating zero-inflation. The probability of belonging to the "nonsusceptible" population, p_i, and the Poisson mean, λ_i, are linked to the explanatory variables through the logit and log links as

$$\operatorname{logit}(p_i) = \mathbf{z}_i'\boldsymbol{\gamma} \text{ and} \tag{2}$$

$$\log(\lambda_i) = \mathbf{x}_i'\boldsymbol{\beta}, \tag{3}$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are vectors of covariates for the ith subject ($i = 1, \ldots, n$) corresponding to the count and zero models, respectively, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_q)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ are the corresponding regression coefficients including the intercepts.

Table 1: The AMAZonn data-adaptive weights. $\widehat{\beta}_{\text{ML}}$ and $\widehat{\gamma}_{\text{ML}}$ denote the ML estimates based on the unpenalized ZIP model, corresponding to count and zero submodels, respectively. SE denotes the standard errors of the corresponding ML estimates.

Weighting Scheme	Count	Zero
AMAZonn - EM AL	$\dfrac{1}{\lvert\widehat{\beta}_{j\text{ML}}\rvert}$	$\dfrac{1}{\lvert\widehat{\gamma}_{j\text{ML}}\rvert}$
AMAZonn - EM SEAL	$\dfrac{SE(\widehat{\beta}_{j\text{ML}})}{\lvert\widehat{\beta}_{j\text{ML}}\rvert}$	$\dfrac{SE(\widehat{\gamma}_{j\text{ML}})}{\lvert\widehat{\gamma}_{j\text{ML}}\rvert}$

For n independent observations, the ZIP log-likelihood function can be written as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = {} & \sum_{y_i=0} \log\left\{e^{z_i'\boldsymbol{\gamma}} + e^{-e^{x_i'\boldsymbol{\beta}}}\right\} \\ & + \sum_{y_i>0}\left\{y_i x_i'\boldsymbol{\beta} + e^{-x_i'\boldsymbol{\beta}}\right\} - \sum_{i=1}^{n}\log\left\{1 + e^{z_i'\boldsymbol{\gamma}}\right\} \\ & - \sum_{y_i>0}\log(y_i!). \end{aligned} \tag{4}$$

2.2. The AMAZonn Method. AMAZonn considers two data-adaptive weights in the EM adaptive LASSO framework: (i) the inverse of the maximum likelihood (ML) estimates (EM AL) and (ii) inverse of the ML estimates divided by their standard errors (EM SEAL). As defined by Tang et al. [6], the EM adaptive LASSO formulation for ZIP regression is given by

$$\widehat{\boldsymbol{\theta}}^{*} = \arg\min\{-L(\boldsymbol{\theta})\} + \nu_1\sum_{j=1}^{p} w_{1j}\left|\beta_j\right| + \nu_2\sum_{j=1}^{p} w_{2j}\left|\gamma_j\right|, \tag{5}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$ is the parameter vector of interest with known weights $w_1 = (w_{11}, \ldots, w_{1p})'$ and $w_2 = (w_{21}, \ldots, w_{2p})'$. As noted by Qian and Yang [7], the inverse of the maximum likelihood (ML) estimates as weights may not always be stable, especially when the multicollinearity of the design matrix is a concern. In order to adjust for this instability, AMAZonn additionally considers the inverse of the ML estimates divided by their standard errors as weights. We refer to these two methods as AMAZonn - EM AL and AMAZonn - EM SEAL, respectively (Table 1).

2.3. The EM Algorithm. In order to efficiently estimate the parameters in the above optimization problem (5), we resort to the EM algorithm. To this end, we define a set of latent variables z_i as follows:

$z_i = 1$ if y_i is from the zero state, and

$$z_i = 0 \text{ if } y_i \text{ is from the count state,} \quad i = 1, \dots, n. \tag{6}$$

We consider the latent variables z_i's as the "missing data" and rewrite the complete-data log-likelihood function in (4) as follows:

$$\begin{aligned} L(\boldsymbol{\theta}) = \sum_{i=1}^{n} [&z_i X_i \gamma - \log(1 + \exp(X_i \gamma)) \\ &+ (1 - z_i)\{y_i X_i \beta - (y_i + 1)\log(1 + X_i \beta)\}]. \end{aligned} \tag{7}$$

With the above formulation, the objective function in (5) can be rewritten as

$$Q^*(\boldsymbol{\theta}) = -L(\boldsymbol{\theta}) + \nu_1 \sum_{j=1}^{p} w_{1j} |\beta_j| + \nu_2 \sum_{j=1}^{p} w_{2j} |\gamma_j|, \tag{8}$$

which can be iteratively solved as follows:

(1) At iteration t, the **E step** computes the expectation of $Q^*(\boldsymbol{\theta})$ by substituting z_i with its conditional expectation given observed data and current parameter estimates

$$\hat{z}_i^{(t)} = \begin{cases} \left(1 + \left[\dfrac{\exp\left(-X_i \hat{\boldsymbol{\gamma}}^{(t)}\right)}{1 + \exp\left(-X_i \hat{\boldsymbol{\beta}}^{(t)}\right)}\right]\right) & \text{if } y_i = 0, \\ 0 & \text{if } y_i > 0. \end{cases} \tag{9}$$

(2) In the **M step**, the expected penalized complete-data log-likelihood (5) can be minimized the with respect to $\boldsymbol{\theta}$ as

$$\begin{aligned} Q^*\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) = -2E(L\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) + \nu_1 \sum_{j=1}^{p} w_{1j} |\beta_j| \\ + \nu_2 \sum_{j=1}^{p} w_{2j} |\gamma_j|. \end{aligned} \tag{10}$$

(3) Continue this process until convergence, $t = 1, 2, \dots$.

It is to be noted that (10) can be further decomposed as

$$Q^*\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) = Q_1^*\left(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)}\right) + Q_2^*\left(\boldsymbol{\gamma} \mid \boldsymbol{\theta}^{(t)}\right), \tag{11}$$

where Q_1^* is the weighted penalized Poisson log-likelihood defined as

$$\begin{aligned} Q_1^*\left(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)}\right) = -2\Bigg[\sum_{i=1}^{n} \left(1 - \hat{z}_i^{(t)}\right) \\ \cdot \{y_i X_i \boldsymbol{\beta} - (y_i + 1)\log(1 + X_i \boldsymbol{\beta})\}\Bigg] \\ + \nu_1 \sum_{j=1}^{p} w_{1j} |\beta_j|, \end{aligned} \tag{12}$$

and Q_2^* is the penalized logistic log-likelihood defined as

$$\begin{aligned} Q_2^*\left(\boldsymbol{\gamma} \mid \boldsymbol{\theta}^{(t)}\right) = -2\left[\sum_{i=1}^{n} \hat{z}_i^{(t)} X_i \boldsymbol{\gamma} - \log(1 + \exp(X_i \boldsymbol{\gamma}))\right] \\ + \nu_2 \sum_{j=1}^{p} w_{2j} |\gamma_j|, \end{aligned} \tag{13}$$

both of which can be minimized separately using computationally efficient coordinate descent algorithms developed for GLMs [9].

2.4. Selection of Tuning Parameters. We select the tuning parameters based on the minimum BIC [10] criterion, which is known to provide better variable selection performance as compared to other information criteria [11]. This can be effortlessly incorporated in our formulation by using existing implementations for zero-inflated count models [3, 4, 6].

3. Oracle Properties

Recently, Tang et al. [6] showed that the EM adaptive LASSO (i.e., AMAZonn - EM AL) enjoys the so-called oracle properties, i.e., the estimator is able to identify the true model consistently, and the resulting estimator is as efficient as *oracle*. Here we extend these results to the AMAZonn - EM SEAL estimator and show that the AMAZonn - EM SEAL estimator also maintains the same theoretical properties. For the sake of completeness, we provide a combined general proof for both AMAZonn estimators.

Without being too rigorous mathematically, recall that the log-likelihood function for the ZIP regression model is given by

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{v}_i) = \sum_{y_i = 0} \log[\psi_i + (1 - \psi_i) f(0; \lambda_i)] \\ + \sum_{y_i > 0} \log[(1 - \psi_i) f(y_i; \lambda_i)], \end{aligned} \tag{14}$$

where $\boldsymbol{v}_i$'s are the observed data (i.i.d observations from the ZIP distribution), $f(.; \lambda_i)$ is the probability mass function of Poisson distribution with parameter $\lambda_i = \exp(X_i \boldsymbol{\beta})$ and $\psi_i = \exp(X_i \boldsymbol{\gamma})/(1 + \exp(X_i \boldsymbol{\gamma}))$, $i = 1, \dots, n$. The corresponding penalized log-likelihood is given by

$$Q(\boldsymbol{\theta}) = -L(\boldsymbol{\theta}; \boldsymbol{v}_i) + \nu_{1n} \sum_{j=1}^{p} w_{1j} |\beta_j| + \nu_{2n} \sum_{j=1}^{p} w_{2j} |\gamma_j|. \tag{15}$$

Let us denote the true coefficient vector as $\boldsymbol{\theta}_{\mathbf{0}} = (\boldsymbol{\beta}_{\mathbf{0}}^T, \boldsymbol{\gamma}_{\mathbf{0}}^T)^T$. Decompose $\boldsymbol{\theta}_{\mathbf{0}} = (\boldsymbol{\theta}_{\mathbf{10}}^T, \boldsymbol{\theta}_{\mathbf{20}}^T)^T$ and assume that $\boldsymbol{\theta}_{20}^T$ contains all zero coefficients. Let us denote the subset of true nonzero coefficients as $\mathscr{A} = \{j : \theta_{j0} \neq 0\}$ and the subset of selected nonzero coefficients as $\widehat{\mathscr{A}} = \{j : \hat{\theta}_j \neq 0\}$. With this formulation, the Fisher information matrix can be written as

$$I(\boldsymbol{\theta}_{\mathbf{0}}) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}, \tag{16}$$

where I_{11} is the Fisher information corresponding the true nonzero submodel. The oracle property of AMAZonn may be developed based on certain mild regularity conditions which are as follows:

(**A1**): The Fisher information matrix $I(\boldsymbol{\theta})$ is finite and positive definite for all values of $\boldsymbol{\theta}$.

(**A2**): There exists functions G_{jkl} such that

$$\frac{\partial^3 L(\boldsymbol{\theta}; \boldsymbol{v}_i)}{\partial\theta_j \partial\theta_k \partial\theta_l} \le G_{jkl}(\boldsymbol{v}_i) \quad \forall\boldsymbol{\theta}, \tag{17}$$

where $g_{jkl} = E_{\boldsymbol{\theta}_0}(G_{jkl}(\boldsymbol{v}_i)) < \infty$ for all j, k, l.

Theorem 1. *Under (A1) and (A2), if $\nu_{1n} \longrightarrow \infty$, $\nu_{2n} \longrightarrow \infty$, $\nu_{1n}/\sqrt{n} \longrightarrow 0$, $\nu_{2n}/\sqrt{n} \longrightarrow 0$, then the AMAZonn estimators obey the following oracle properties:*

(1) *consistency in variable selection:* $\lim_n P(\widehat{\mathscr{A}} = \mathscr{A}) = 1$, *and*

(2) *asymptotic normality of the nonzero coefficients:* $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathbf{0}}) \longrightarrow_d \mathscr{N}(\mathbf{0}, I_{11}^{-1})$.

4. Simulation Studies

In this section, we conduct simulation studies to evaluate the finite sample performance of AMAZonn. For comparison purposes, the performance of both AMAZonn and EM LASSO is evaluated. For each simulated dataset, the associated tuning parameters are selected by the minimum BIC criterion for all the methods under consideration. All the examples reported in this section are obtained from published papers with slight modifications within the scope of the current study [11, 12].

Specially, three scenarios are considered: in the data generating models of Simulations 1 and 2, we consider all continuous predictors, whereas in Simulation 3, both continuous and categorical variables are included. For each experimental instance, we randomly partition the data into training and test sets: models are fitted on the training set and prediction errors based on mean absolute scaled error (MASE) are calculated on the held-out samples in the test set. For an exhaustive comparison, we considered three sets of sample sizes $\{n_T, n_P\} = \{200, 200\}, \{500, 500\}$, and $\{1000, 1000\}$, where n_T and n_P represent the size of the training and test data, respectively. The corresponding regression coefficients and intercepts are chosen so that a desired level of sparsity proportion (ϕ) is achieved. In order to remain as model-agnostic as possible, we consider the same set of predictors for both zero and count submodels (i.e., $\mathbf{X} = \mathbf{Z}$). Such models are common in many practical applications where no domain-specific prior information about the zero-inflation mechanism is available. Below we provide the detailed data generation steps for both simulation examples:

Simulation 1.

(1) Generate 40 predictors from the multivariate normal distribution with mean vector $\mathbf{0}$, variance vector $\mathbf{1}$, and variance-covariance matrix V, where the elements of V are $\rho^{|j_1 - j_2|}$ $\forall j_1 \neq j_2 = 1, \ldots, 40$. The values of pairwise correlation ρ varies from 0 (uncorrelated) to 0.4 (moderate collinearity) to 0.8 (high collinearity).

(2) The count and zero regression parameters are chosen as follows:

$$\begin{aligned}
&(\beta_1, \ldots, \beta_8) \\
&\quad = (-1, -0.5, -0.25, -0.1, 0.1, 0.25, 0.5, 0.75)', \\
&(\beta_9, \ldots, \beta_{16}) = (0.2, \ldots, 0.2)', \\
&(\beta_{17}, \ldots, \beta_{40}) = (0, \ldots, 0)', \\
&(\gamma_1, \ldots, \gamma_8) \\
&\quad = (-0.4, -0.3, -0.2, -0.1, 0.1, 0.2, 0.3, 0.4)', \\
&(\gamma_9, \ldots, \gamma_{16}) = (0.2, \ldots, 0.2)', \\
&(\gamma_{17}, \ldots, \gamma_{40}) = (0, \ldots, 0)'.
\end{aligned} \tag{18}$$

(3) The zero-inflated count outcome y is simulated according to (1) with the above parameters and input data.

Simulation 2. It is similar to Simulation 1 except that the count and zero regression parameters are chosen as follows:

$$\begin{aligned}
&(\beta_1, \ldots, \beta_{10}) = (0.05, -0.25, 0.05, 0.25, \\
&\quad -0.15, 0.15, 0.25, -0.2, 0.25, -0.25)', \\
&(\beta_{11}, \ldots, \beta_{30}) = (-0.2, 0.25, 0.15, \\
&\quad -0.25, 0.2, 0, \ldots, 0)', \\
&(\beta_{31}, \ldots, \beta_{40}) = (0.27, -0.27, 0.14, 0.2, \\
&\quad -0.2, 0.2, 0, \ldots, 0)', \\
&(\gamma_1, \ldots, \gamma_{10}) = (-0.5, -0.4, -0.3, -0.2, \\
&\quad -0.1, 0.1, 0.2, 0.3, 0.4, 0.5)', \\
&(\gamma_{11}, \ldots, \gamma_{30}) = (-0.2, 0.25, 0.15, -0.25, 0.2, 0, \ldots, 0)', \\
&(\gamma_{31}, \ldots, \gamma_{40}) = (0.27, -0.27, -0.14, -0.2, \\
&\quad -0.2, 0.2, 0, \ldots, 0)'.
\end{aligned} \tag{19}$$

Simulation 3.

(1) First simulate $X_1, \ldots, X_6$ independently from the standard normal distribution. Consider the following as the continuous predictors: $\{X_1\}, \{X_2\}, \{X_3, X_3^2, X_3^3\}, \{X_4\}, \{X_5\}$ and $\{X_6, X_6^2, X_6^3\}$.

(2) Simulate 5 continuous variables from the multivariate normal distribution with mean 0, variance 1, and AR(ρ) correlation structure for varying ρ in $\{0, 0.4,$

Table 2: Results of Simulations 1–3. Average (over 200 replications) of Mean Absolute Scale Errors (MASEs) of AMAZonn and EM LASSO is reported.

			Simulation 1			Simulation 2			Simulation 3		
ρ	ϕ	n	AMAZonn - EM SEAL	AMAZonn - EM AL	EM LASSO	AMAZonn - EM SEAL	AMAZonn - EM AL	EM LASSO	AMAZonn - EM SEAL	AMAZonn - EM AL	EM LASSO
0.0	0.3	200	0.91	0.92	0.91	0.60	0.61	0.62	0.97	1.03	1.00
		500	0.90	0.90	0.91	0.60	0.60	0.61	0.97	0.99	1.00
		1000	0.91	0.91	0.92	0.58	0.58	0.60	0.97	0.98	0.98
	0.4	200	1.12	1.13	1.12	0.75	0.75	0.76	1.18	1.23	1.23
		500	1.05	1.05	1.06	0.73	0.73	0.74	1.11	1.17	1.20
		1000	1.03	1.03	1.04	0.71	0.71	0.72	1.11	1.16	1.16
	0.5	200	1.28	1.28	1.27	0.87	0.87	0.87	1.40	1.46	1.46
		500	1.16	1.16	1.17	0.84	0.84	0.85	1.28	1.33	1.36
		1000	1.15	1.15	1.19	0.80	0.80	0.82	1.23	1.30	1.31
0.4	0.3	200	1.05	1.06	1.09	0.63	0.63	0.63	0.96	1.01	0.99
		500	1.04	1.04	1.05	0.61	0.61	0.62	0.95	0.97	0.99
		1000	0.96	0.96	0.98	0.58	0.58	0.59	0.97	0.98	0.98
	0.4	200	1.21	1.22	1.22	0.75	0.75	0.76	1.19	1.22	1.23
		500	1.18	1.18	1.21	0.71	0.71	0.72	1.14	1.19	1.22
		1000	1.13	1.14	1.18	0.68	0.68	0.70	1.13	1.18	1.17
	0.5	200	1.42	1.43	1.42	0.83	0.84	0.83	1.34	1.40	1.43
		500	1.26	1.26	1.32	0.80	0.81	0.82	1.27	1.32	1.35
		1000	1.23	1.23	1.30	0.75	0.75	0.77	1.27	1.34	1.33
0.8	0.3	200	1.32	1.31	1.36	0.62	0.63	0.63	0.96	1.00	1.01
		500	1.13	1.13	1.23	0.59	0.59	0.61	0.97	0.99	1.01
		1000	1.13	1.13	1.21	0.56	0.56	0.58	0.95	0.96	0.96
	0.4	200	1.52	1.52	1.58	0.71	0.72	0.72	1.18	1.21	1.23
		500	1.31	1.32	1.45	0.68	0.68	0.69	1.12	1.19	1.20
		1000	1.24	1.24	1.37	0.64	0.64	0.64	1.12	1.17	1.16
	0.5	200	1.56	1.58	1.61	0.78	0.78	0.78	1.37	1.42	1.44
		500	1.44	1.45	1.65	0.73	0.73	0.76	1.29	1.34	1.39
		1000	1.33	1.36	1.52	0.69	0.70	0.69	1.26	1.33	1.34

0.8} as before, and quantile-discretize each of them into 5 new variables based on their quantiles: $(-\infty, \Phi^{-1}(1/5)]$, $(\Phi^{-1}(1/5), \Phi^{-1}(2/5)]$, $(\Phi^{-1}(2/5), \Phi^{-1}(3/5)]$, $(\Phi^{-1}(3/5), \Phi^{-1}(4/5)]$, and $(\Phi^{-1}(4/5), \infty)$, leading to a total of 20 categorical variables.

(3) With the above input data and parameters, the zero-inflated count outcome y is simulated according to (1), where the two sets of regression parameters are chosen as follows:

$$\begin{aligned}
(\beta_1, \ldots, \beta_{10}) &= \left(0, 0, 0.1, 0.2, 0.1, 0, 0, \frac{2}{3}, -1, \frac{1}{3}\right), \\
(\beta_{11}, \ldots, \beta_{30}) &= (-2, -1, 1, 2, 0, \ldots, 0), \\
(\gamma_1, \ldots, \gamma_{10}) &= \left(0, 0, 0.1, 0.2, 0.1, 0, 0, \frac{2}{3}, -1, \frac{1}{3}\right), \\
(\gamma_{11}, \ldots, \gamma_{30}) &= (-2, -1, 1, 2, 0, \ldots, 0).
\end{aligned} \tag{20}$$

The resulting performance measures iterated over 200 replications (Table 2) reveal that AMAZonn performs as well as or better than EM LASSO in most of the simulation scenarios. For highly collinear designs, AMAZonn - EM SEAL stands out to be the best estimator for almost every sample size and zero-inflation proportion, highlighting the benefit of incorporating data-adaptive weights based on both ML estimates and their standard errors. This phenomenon is also apparent in the analysis of German health care data in Section 5, where the parameter estimates from the AMAZonn - EM SEAL method appear to be more parsimonious than those from other methods.

5. Application to German Health Care Demand Data

Next, we apply our method to the German health care demand data [3], a subset of the German Socioeconomic Panel (GSOEP) dataset [13], which has also been used for

Table 3: Summary of predictors in German health care demand data.

Variables	Mean (sd) or Frequency	Description
health	6.84 (2.19)	health satisfaction: 0 (low) - 10 (high)
handicap	216 / 1596	1 : handicap, 0 : otherwise
hdegree	6.16 (18.49)	degree of handicap in percentage points
married	1257 / 555	1 : married, 0 : otherwise
schooling	11.83 (2.49)	years of schooling
hhincome	4.52 (2.13)	household income per month in German marks/1000
children	703 / 1109	1 : children under 16 in household, 0 : otherwise
self	153 / 1659	1 : self-employed, 0 : otherwise
civil	198 / 1614	1 : civil servant, 0 : otherwise
bluec	566 / 1246	1 : blue collar employee, 0 : otherwise
employed	1506 / 306	1 : employed, 0 : otherwise
public	1535 / 277	1 : public health insurance, 0 : otherwise
addon	33 / 1779	1 : addon insurance, 0 : otherwise
age30	1480 / 332	1 if age ≥ 30
age35	1176 / 636	1 if age ≥ 35
age40	919 / 893	1 if age ≥ 40
age45	716 / 1096	1 if age ≥ 45
age50	535 / 1227	1 if age ≥ 50
age55	351 / 1461	1 if age ≥ 55
age60	147 / 1665	1 if age ≥ 60

Table 4: Model selection performance of EM LASSO and AMAZonn on German health care data.

Methods	BIC	Time (in seconds)
EM LASSO	9062.744	50.252
AMAZonn - EM AL	9002.487	26.215
AMAZonn - EM SEAL	**8982.924**	26.528

illustration purposes in previous studies [3, 14]. The original data contains number of doctor office visits for 1, 812 West German men aged 25 to 65 years in the last three months of 1994 (response variable of interest), which is supplemented with complementary information on twelve annual waves from 1984 to 1995 including health care utilization, current employment status, and insurance arrangements under which subjects are protected [3]. The goal of the original study was to investigate how the employment characteristics of the German nationals are related to their health care demand. The distribution of the dependent variable (Figure 1) reveals that many doctor visits are zeros (41.2%), confirming that classical methods such as Poisson regression are inappropriate for modeling this outcome.

In the model fitting process, along with the original variables, the interactions between age groups and health condition are also considered, resulting in 28 candidate predictors (Table 3). The fitting results from the full models indicate that both EM adaptive LASSO methods provide competitive model selection performance (Table 4), often leading to sparser model selection than EM LASSO (Table 5). In addition, the AMAZonn - EM SEAL method appears to choose even fewer numbers of variables. Such feature of AMAZonn - EM SEAL can be appealing in many practical situations, where data collinearity between variables is a concern and a more aggressive feature selection is desired. While the computational overheads of both EM adaptive LASSO methods are similar, they are an order of magnitude faster than EM LASSO (Table 4), further confirming that AMAZonn offers a viable alternative to existing methods.

6. Discussion

In recent years, there has been a huge influx of zero-inflated count measurements spanning several disciplines including biology, public health, and medicine. This has motivated the widespread use of zero-inflated count models in many practical applications such as metagenomics, single-cell RNA sequencing, and health care research. In this article, we propose the AMAZonn method for adaptive variable selection in ZIP regression models. Both our simulation and real data experience suggest that AMAZonn can outperform EM LASSO under a variety of regression settings while maintaining the desired theoretical properties and computational convenience. Our preliminary results are rather encouraging, and for practical purposes, we provide a publicly available R package implementing this method: https://github.com/himelmallick/AMAZonn.

We envision a number of improvements that may further refine AMAZonn's performance. While AMAZonn relies on ML estimates to construct the weight vector, these estimates may not be available in ultrahigh dimensions [7]. Alternative initialization schemes could further improve on this such as the ridge estimates [15]. Extension to other zero-inflated models such as marginalized zero-inflated count regression [16, 17], two-part and hurdle models [18], and multiple-inflation models [19] can form a useful basis for further

Table 5: Estimated coefficients from the best-fitting ZIP models in German health care demand data analysis.

(a)

Methods								Count Coefficients						
	(Intercept)	hlth	handicap	ddegree	married	schooling	hhincome	children	self	civil	bluec	employed	public	addon
EM LASSO	2.322	-0.14	0.207	-0.002	-0.97	0.0	0.0	0.078	-0.178	-0.166	0.038	-0.106	0.089	0.205
AMAZonn - EM AL	2.305	-0.135	0.111	0.0	-0.947	0.0	0.0	0.079	-0.234	-0.245	0.0	-0.059	0.043	0.205
AMAZonn - EM SEAL	2.378	-0.142	0.098	0.0	-0.066	0.0	0.0	0.046	-0.189	-0.222	0.0	-0.055	0.0	0.14
Methods								Count Coefficients						
	ag30	ag35	ag40	ag45	ag50	ag55	ag60	ag30:hlth	ag35:hlth	ag40:hlth	ag45:hlth	ag50:hlth	ag55:hlth	ag60:hlth
EM LASSO	0.0	0.0	0.0	0.586	0.0	-0.27	0.081	0.0	0.0	-0.006	-0.076	0.006	0.082	-0.034
AMAZonn - EM AL	0.0	0.0	-0.047	0.769	0.0	-0.402	0.099	0.0	0.0	0.0	-0.101	0.0	0.106	-0.034
AMAZonn - EM SEAL	0.0	0.0	0.0	0.586	0.0	-0.25	0.0	0.0	0.0	0.0	-0.081	0.0	0.081	-0.017

(b)

Methods								Zero Coefficients						
	(Intercept)	hlth	handicap	ddegree	married	schooling	hhincome	children	self	civil	bluec	employed	public	addon
EM LASSO	-2.193	-0.262	-0.098	-0.003	-0.121	0.0	-0.012	0.253	0.112	0.134	0.0	0.0	-0.012	0.0
AMAZonn - EM AL	-2.226	-0.261	-0.162	0.0	0.0	0.0	0.0	0.163	0.0	0.0	0.0	0.0	0.0	0.0
AMAZonn - EM SEAL	-2.403	-0.283	0.0	0.0	-0.053	0.0	0.0	0.238	0.0	0.0	0.0	0.0	0.0	0.0
Methods								Zero Coefficients						
	ag30	ag35	ag40	ag45	ag50	ag55	ag60	ag30:hlth	ag35:hlth	ag40:hlth	ag45:hlth	ag50:hlth	ag55:hlth	ag60:hlth
EM LASSO	0.0	0.0	0.0	0.0	-0.459	0.0	-0.217	0.013	0.0	0.005	0.0	0.0	0.023	0.0
AMAZonn - EM AL	0.047	0.0	0.065	0.009	-0.527	0.0	-0.198	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AMAZonn - EM SEAL	0.0	0.0	0.0	0.0	-0.443	0.0	0.0	0.009	0.0	0.0	0.0	0.0	0.0	0.0

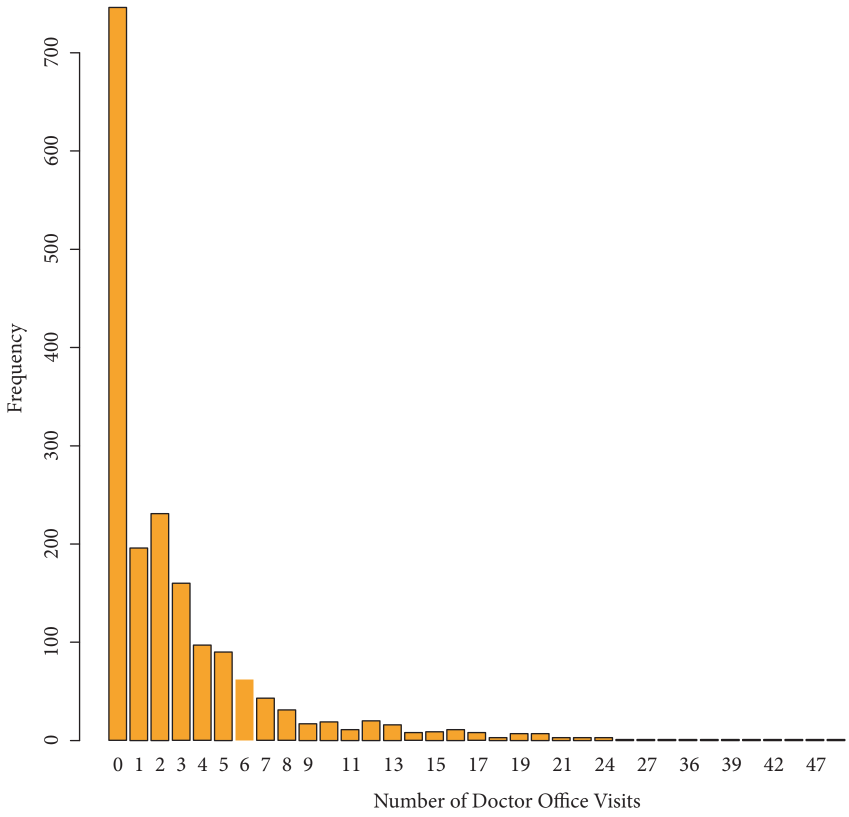


Figure 1: Number of doctor office visits in the German health care data.

investigations. Although we only focused on variable selection for fixed effects models, future work could include an extension to other regularization problems such as grouped variable selection [12, 20] as well as sparse mixed effects models [21].

Appendix

Proof. It is to be noted that both logistic and Poisson distributions belong to the exponential family. Since the objective function in (10) can be decomposed into weighted logistic and Poisson log-likelihoods (each belonging to the GLM family without the penalties), Theorem 1 is the direct application of Theorem 4 in Zou [22]. Therefore, if $\nu_{1n} \longrightarrow \infty$, $\nu_{2n} \longrightarrow \infty$, $\nu_{1n}/\sqrt{n} \longrightarrow 0$, and $\nu_{2n}/\sqrt{n} \longrightarrow 0$, then both the AMAZonn - EM AL and AMAZonn - EM SEAL estimators hold the oracle properties: with probability tending to 1, the estimate of zero coefficients is 0, and the estimate for nonzero coefficients has an asymptotic normal distribution with mean being the true value and variance which approximately equals the submatrix of the Fisher information matrix containing nonzero coefficients. Hence the proof is complete. □

Data Availability

The German Healthcare dataset used in the paper is publicly available from others (https://cran.r-project.org/web/packages/HDtweedie/index.html) and the software is publicly available at https://github.com/himelmallick/AMAZonn.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Prithish Banerjee, Broti Garai, and Himel Mallick contributed equally to this work.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the manuscript. This work was supported in part by the research computing resources acquired and managed by University of Alabama at Birmingham IT Research Computing. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the University of Alabama at Birmingham.

References

[1] W. H. Greene, *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models*, New York University, New York, NY, 1994.

[2] D. Lambert, "Zero-inflated poisson regression, with an application to defects in manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.

[3] Z. Wang, S. Ma, and C.-Y. Wang, “Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany,” *Biometrical Journal*, vol. 57, no. 5, pp. 867–884, 2015.

[4] Z. Wang, S. Ma, C.-Y. Wang, M. Zappitelli, P. Devarajan, and C. Parikh, “EM for regularized zero-inflated regression models with applications to postoperative morbidity after cardiac surgery in children,” *Statistics in Medicine*, vol. 33, no. 29, pp. 5192–5208, 2014.

[5] H. Mallick and H. K. Tiwari, “EM adaptive LASSO-a multilocus modeling strategy for detecting SNPs associated with zero-inflated count phenotypes,” *Frontiers in Genetics*, vol. 7, 2016.

[6] Y. Tang, L. Xiang, and Z. Zhu, “Risk Factor Selection in Rate Making: EM Adaptive LASSO for Zero-Inflated Poisson Regression Models,” *Risk Analysis*, vol. 34, no. 6, pp. 1112–1127, 2014.

[7] W. Qian and Y. Yang, “Model selection via standard error adjusted adaptive lasso,” *Annals of the Institute of Statistical Mathematics*, vol. 65, no. 2, pp. 295–318, 2013.

[8] Z. Y. Algamal and M. H. Lee, “Adjusted Adaptive LASSO in High-dimensional Poisson Regression Model,” *Modern Applied Science (MAS)*, vol. 9, no. 4, 2014.

[9] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate descent,” *Journal of Statistical Software* , vol. 33, no. 1, pp. 1–22, 2010.

[10] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[11] J. Huang, S. Ma, H. Xie, and C.-H. Zhang, “A group bridge approach for variable selection,” *Biometrika*, vol. 96, no. 2, pp. 339–355, 2009.

[12] S. Chatterjee, S. Chowdhury, H. Mallick, P. Banerjee, and B. Garai, “Group regularization for zero-inflated negative binomial regression models with an application to health care demand in Germany,” *Statistics in Medicine*, vol. 37, no. 20, pp. 3012–3026, 2018.

[13] R. T. Riphahn, A. Wambach, and A. Million, “Incentive effects in the demand for health care: A bivariate panel count data estimation,” *Journal of Applied Econometrics*, vol. 18, no. 4, pp. 387–405, 2003.

[14] M. Jochmann, “What belongs where? Variable selection for zero-inflated count models with an application to the demand for health care,” *Computational Statistics*, vol. 28, no. 5, pp. 1947–1964, 2013.

[15] A. E. Hoerl and R. W. Kennard, “Ridge regression: biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[16] D. L. Long, J. S. Preisser, A. H. Herring, and C. E. Golin, “A marginalized zero-inflated Poisson regression model with overall exposure effects,” *Statistics in Medicine*, vol. 33, no. 29, pp. 5151–5165, 2014.

[17] V. A. Smith and J. S. Preisser, “Direct and flexible marginal inference for semicontinuous data,” *Statistical Methods in Medical Research*, vol. 26, no. 6, pp. 2962–2965, 2016.

[18] V. A. Smith, B. Neelon, J. S. Preisser, and M. L. Maciejewski, “A marginalized two-part model for longitudinal semicontinuous data,” *Statistical Methods in Medical Research*, vol. 26, no. 4, pp. 1949–1968, 2017.

[19] X. Su, J. Fan, R. A. Levine, X. Tan, and A. Tripathi, “Multiple-inflation Poisson model with L_1 regularization,” *Statistica Sinica*, vol. 23, no. 3, pp. 1071–1090, 2013.

[20] S. Chowdhury, S. Chatterjee, H. Mallick, H. Banerjee, and B. Garai, “Group regularization for zero-inflated poisson regression models with an application to insurance ratemaking,” *Journal of Applied Statistics*, 2018, In Press.

[21] A. Groll and G. Tutz, “Variable selection for generalized linear mixed models by L_1-penalized estimation,” *Statistics and Computing*, vol. 24, no. 2, pp. 137–154, 2014.

[22] H. Zou, “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

Hindawi
Journal of Probability and Statistics
Volume 2018, Article ID 3506794, 12 pages
https://doi.org/10.1155/2018/3506794

Research Article

Detecting Spatial Clusters via a Mixture of Dirichlet Processes

Meredith A. Ray,[1] **Jian Kang**,[2] **and Hongmei Zhang**[3]

[1] *Division of Epidemiology, Biostatistics, and Environmental Health, University of Memphis, 220 Robison Hall, Memphis, TN 38152, USA*
[2] *Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA*
[3] *Division of Epidemiology, Biostatistics, and Environmental Health, University of Memphis, 224 Robison Hall, Memphis, TN 38152, USA*

Correspondence should be addressed to Meredith A. Ray; maray@memphis.edu

Received 14 May 2018; Revised 14 November 2018; Accepted 25 November 2018; Published 18 December 2018

Guest Editor: Yichuan Zhao

We proposed an approach that has the ability to detect spatial clusters with skewed or irregular distributions. A mixture of Dirichlet processes (DP) was used to describe spatial distribution patterns. The effects of different batches of data collection efforts were also modeled with a Dirichlet process. To cluster spatial foci, a birth-death process was applied due to its advantage of easier jumping between different numbers of clusters. Inferences of parameters including clustering were drawn under a Bayesian framework. Simulations were used to demonstrate and assess the method. We applied the method to an fMRI meta-analysis dataset to identify clusters of foci corresponding to different emotions.

1. Introduction

This work was motivated by a study aiming to detect centers of activated foci from a function magnetic resonance imaging (fMRI) metadataset. In summary, fMRI metadata is a collection of fMRI studies identifying areas of the brain that are significantly activated by stimuli to examine a specific outcome. FMRIs are expensive which leads to small sample sizes and therefore can use metadata to increase sample size and power. To identify spatial clusters, finite mixture models are generally implemented [1–3]. Mixture components, representing a cluster, typically share a common parametric family with each component containing respective parameters [1, 2]. Each component also has a mixing proportion or weight that is respective to the frequency of data in that component [1]. Because of the model's ease of implementation, this allows various applications such as pattern recognition, computer vision, signal and image analysis, and machine learning to list a few [4].

One commonly used distribution in the aforementioned finite mixtures is the normal distribution [1, 5], appreciating its established properties and, in the Bayesian context, conjugacy. However, when it comes to clustering, assuming normality for each part of the mixture can potentially lead to oversensibility, e.g., when one cluster is formed by a mixture of two normal distributions but with rather close centers. This type of oversensibility in many research fields should be avoided; one example is emotional foci inferred from brain imaging data, where a certain emotion is covered by a wide region.

To infer the number of clusters, under the Bayesian framework, different methods have been proposed. Reversible jump Markov chain Monte Carlo has been commonly used to infer the number of clusters [6, 7]. At each iteration, a decision is made between splitting one cluster to two, combining two clusters to one, or no movement. One potential difficulty of this approach is the risk of being trapped at a local maximum. Recently, the Dirichlet process (DP) has been used often to estimate the number of clusters [8, 9]. This process has the ability to capture irregular patterns. DP has the ability to detect clusters without being burdened about additional clustering parameters. However, this feature of clustering also has an inherit weakness in that it tends to produce more clusters, making interpretations more difficult.

To overcome the aforementioned gaps, we implemented a mixture of Dirichlet processes (DP). These processes have the ability to describe irregular patterns [8, 9] and by using it as our common parametric family allows the model to identify more complex patterns than the normal distribution would be able to identify. Furthermore, motivated from our previous work with the spatial Cox process application in [10], we elected to incorporate the birth-death process to statistically determine the number of clusters. Compared to other clustering approaches noted earlier, the birth-death process has the advantage of quick convergence and, by controlling birth rate, embraces a potential of generating redundant clusters.

The remainder of the article is organized as follows. In Section 2, we introduce the model structure, notation, and priors and hyperpriors, simulations are presented in Section 3, the application of the model to an fMRI meta-analysis dataset is in Section 4, and conclusions and discussion are in Section 5.

2. Methods

2.1. The Model. We let s_{rj} denote the (x, y, z) coordinate of a spatial point in a three-dimensional space, in particular, point j in group (study) r, $j = 1, \ldots, J_r$ and $r = 1, \ldots, R$. We have $\sum_{r=1}^{R} J_r = n$, where n is the total number of observed points. It straightforwardly follows that $\boldsymbol{s} = (s_{1,1} \cdots s_{R,J_R})$ represents all points in the study. We model s_{rj} as

$$s_{rj} = p_r + \theta_{rj} + \epsilon, \tag{1}$$

where p_r denotes the effect of group r, while θ_{rj} represents the mean of s_{rj} for the j^{th} point in group r after adjusting for group effects, p_r, and ϵ denotes some random error. By modeling the random error as a standard multivariate normal distribution, the distribution of s_{rj} satisfies

$$s_{rj} \sim MVN_3\left(p_r + \theta_{rj}, \Sigma\right) \tag{2}$$

with

$$\begin{aligned} P\left(s_{rj} \mid p_r, \theta_{rj}, \Sigma\right) &= (2\pi)^{-3/2} \left|\Sigma\right|^{-1/2} \\ &\cdot \exp\left[-\frac{1}{2}\left(s_{rj} - p_r - \theta_{rj}\right)' \Sigma^{-1} \left(\left(s_{rj} - p_r - \theta_{rj}\right)\right)\right], \end{aligned} \tag{3}$$

where $\Sigma = \sigma^2 I_3$ is the covariance matrix.

2.2. Prior and Hyperprior Distributions. We start from the prior distribution of θ_{rj}. To detect underlying clusters of s_{rj} due to similarities of θ_{rj}, we describe the prior of θ_{rj} as a mixture of distribution G_k, $k = 1, \ldots, K$. Common choices of G_k are normal distributions. To improve flexibility, we relax such normalizing assumptions in the mixture and assume G_k is generated from a DP, i.e., $G_k \sim DP(\alpha, G_{0k})$, where α is the precision parameter and G_{0k} is a base distribution and taken as $G_{0k} = MVN_3(\boldsymbol{\mu_k}, \Sigma_k)$. In particular, $\theta_{rj} \sim \sum_{k=1}^{K} \pi_k G_k$, where $0 < \pi_k < 1$ such that $\sum_{k=1}^{K} \pi_k = 1$. For the number of clusters K, we assign a truncated Poisson prior distribution to K, $P(K) = (\lambda^K/K!)\exp(-\lambda)$, $K = 1, \ldots, n$. We assign a Dirichlet distribution with parameter 1 to $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, which implies that $\boldsymbol{\pi}$ is k-dimensional uniformly distributed. The prior distribution of $\boldsymbol{\mu_k}$ from the base distribution G_{0k} is chosen to be $\boldsymbol{\mu_k} \sim MVN_3(\boldsymbol{\xi}, \kappa^{-1})$, where $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)$, which are known and set as the midpoint of observed interval of variation of the data. Parameter κ is set as

$$\kappa = \begin{bmatrix} \dfrac{1}{R_1^2} & 0 & 0 \\ 0 & \dfrac{1}{R_2^2} & 0 \\ 0 & 0 & \dfrac{1}{R_3^2} \end{bmatrix}, \tag{4}$$

where R_1^2, R_2^2, and R_3^2 are the range of the data for each dimension. This prior setting is adopted from [1] and we feel it is reasonable for this setting given the fact that the number and location of clusters are unknown. We let $\Sigma_k = \sigma_k I_3$ with $\sigma_k \sim IG(20, 0.5)$ since the range of the observed data is small. For group effect, p_r, we assume it is small with $p_r \sim G_p$, where $G_p \sim DP(\alpha_p, G_{0p})$ with α_p specified later and $G_{0p} = TMVN_3(\boldsymbol{\mu_p}, \Sigma_p, lower = -l, upper = l)$, where the lower and upper limits are defined as 10% of the absolute range of the data. We let $\boldsymbol{\mu_p} \sim MVN_3(\mathbf{0}, I_3)$ and $\Sigma_p = \sigma_p^2 I_3$, with $\sigma_p^2 \sim IG(5, 0.5)$. The variance component of the random error, $\Sigma = \sigma^2 I_3$, σ^2 is assumed to follow a relatively noninformative Inverse Gamma (IG) distribution, $\sigma^2 \sim IG(0.5, 0.5)$. The precision parameters α and α_p are selected by minimizing the deviance information criterion (DIC) [11, 12].

2.3. Conditional Posterior Distributions and Posterior Computing. Sampling parameter estimates from their posterior distributions can be achieved via Gibbs sampler, in which the statistical inference on the number of clusters is modeled using the birth-death process. The birth-death process is one type of continuous-time Markov chain originally introduced in [13]. This process is often used to simulate realizations of point processes as they can be difficult to directly sample from [1]. These realizations are further used for likelihood inferences for model parameters [1]. The birth-death scheme allows events to randomly occur throughout the chain; these events are either a "birth" or "death." If a birth occurs, the number of components increases by one, while if a death occurs, the number of components decreases by one.

Recall considering a finite mixture prior for θ_{rj} such that all θ_{rj} are assumed independently distributed with each generated from one of K Dirichlet processes denoted as G_k, i.e.,

$$P\left(\theta_{rj} \mid K, \boldsymbol{\pi}, \boldsymbol{\phi}, \alpha\right) = \sum_{k=1}^{K} \pi_k G_k\left(\alpha, G_{0k}\right), \tag{5}$$

where each $G_k(\alpha, G_{0k})$ represents a DP but K is unknown, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ are the mixing proportions, and $\boldsymbol{\phi} = (\boldsymbol{\mu}_1, \Sigma_k, \ldots, \boldsymbol{\mu}_K, \Sigma_K)$ are the component specific parameters for each DP. For cluster assignment, we introduce an index variable Z_{rj} that indicates the assignment of observation rj and Z_{rj} takes the values of 1 to K. Denoted by $z_{rj} \in \boldsymbol{z}$, where $\boldsymbol{z} = (z_{1,1}, \ldots, z_{R,J_R})$ represents the realization of independently and identically distributed discrete random variables $\mathbf{Z} = (Z_{1,1}, \ldots, Z_{R,J_R})$ with probability mass function,

$$P\left(Z_{rj} = k \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \eta\right) = \pi_k \quad (k = 1, \ldots, K). \tag{6}$$

The joint posterior distribution is proportional to

$$\begin{aligned} P(\boldsymbol{\vartheta} \mid \boldsymbol{s}) \propto & \prod_r^R P\left(p_r \mid G_p\right) P\left(G_p \mid \alpha, G_{0p}\right) \\ & \cdot G_{0p}\left(p_r; \mu_p, \sigma_p^2\right) P\left(\mu_p\right) P\left(\sigma_p^2\right) \\ & \cdot \prod_j^{J_r} P\left(s_{rj} \mid \boldsymbol{\phi}\right) \\ & \times P\left(\theta_{rj} \mid \boldsymbol{z}, \boldsymbol{\pi}, G_1, \ldots, G_K\right) \\ & \cdot \prod_{k=1}^{K} P\left(G_k \mid \alpha, G_{0k}\right) G_{0k}\left(\theta_{rj}; \mu_k, \sigma_k^2\right) \\ & \cdot P\left(\mu_k\right) P\left(\sigma_k^2\right) \times P(\boldsymbol{\pi}) P(K) P(\mathbf{Z}) P\left(\sigma^2\right), \end{aligned} \tag{7}$$

where $\boldsymbol{\vartheta} = (p_r, \theta_{rj}, \sigma^2, \alpha_p, \boldsymbol{\mu_p}, \sigma_p^2, K, \boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}, \sigma_1^2, \ldots, \sigma_K^2, \alpha, \boldsymbol{\pi})$ is a vector of all estimable parameters. From here, the birth-death algorithm and Markov chain can be described for introducing and assigning clusters for θ_{rj}:

(1) Starting with the initial model $y = \{(\pi_1, \phi_1), \ldots, (\pi_K, \phi_K)\}$, let (π_k, ϕ_k) represent the mixing proportion and cluster specific parameters for the unique θ_{rj} clusters. Let the birth rate be $\beta(y) = \lambda_b$.

(2) Calculate the death rate for each component:

$$\delta_k(y) = \frac{L(y \setminus (\pi_k, \phi_k))}{L(y)} \frac{P(K-1 \mid \alpha, \cdot)}{KP(K \mid \alpha, \cdot)} \qquad (k = 1, \ldots, K). \tag{8}$$

(3) Calculate the total death rate $\delta(y) = \sum_{k=1}^{K} \delta_k(y)$. To quicken convergence, we elected not to model the time to next jump as exponential and allowed an event to occur at each iteration of the Markov chain.

(4) Simulate the type of event, birth or death, with the respective probabilities:

$$\begin{aligned} P(birth) &= \frac{\beta(y)}{\beta(y) + \delta(y)}, \\ P(death) &= \frac{\delta(y)}{\beta(y) + \delta(y)}. \end{aligned} \tag{9}$$

(5) Adjust the model y to reflect the birth or death by the following:

(i) Birth: Simulate a new component (π_{K+1}, ϕ_{K+1}) from each parameter's respective (independent) prior distributions, π_{K+1} from $K(1-\pi)^{K-1}$ and ϕ_{K+1} from base distribution, such that the model becomes $y = \{(\pi_1, \phi_1), \ldots, (\pi_K, \phi_K), (\pi_{K+1}, \phi_{K+1})\}$. It can be mentioned that $K(1-\pi)^{K-1}$ is the Beta distribution with parameters $(1, K)$ and can be easily simulated from $Y_1 \sim G(1,1)$ and $Y_2 \sim G(K,1)$ such that $\pi_{K+1} = Y_1/(Y_1+Y_2)$.

(ii) Death: Select a component to die with the probabilities $\delta_k(y)/\delta(y)$ for $k = 1, \ldots, K$ such that the model becomes $y = \{(\pi_1, \phi_1), \ldots, (\pi_{K-1}, \phi_{K-1})\}$.

(6) Given the current state of the model at time t, simulate values for all remaining parameters.

(7) Go to step (2).

Incorporating the birth-death process into our model, we need to further define the likelihood for removing cluster i:

$$\begin{aligned} P(\boldsymbol{\Phi} \setminus (\pi_i, G_i) \mid \mathbf{s}) \propto & \prod_r^R P\left(p_r \mid G_p\right) P\left(G_p \mid \alpha, G_{0p}\right) \\ & \cdot G_{0p}\left(p_r; \mu_p, \sigma_p^2\right) P\left(\mu_p\right) P\left(\sigma_p^2\right) \prod_j^{J_r} P\left(s_{rj} \mid \boldsymbol{\phi}\right) \\ & \times P\left(\theta_{rj} \mid \boldsymbol{z}, \boldsymbol{\pi}, G_1, \ldots, G_K\right) \prod_{k=1}^{K^{(i)}} P\left(G_k \mid \alpha, G_{0k}\right) \\ & \cdot G_{0k}\left(\theta_{rj}; \mu_k, \sigma_k^2\right) \times P\left(\mu_k\right) P\left(\sigma_k^2\right) P(\boldsymbol{\pi}) P(K) P(\mathbf{Z}) \\ & \cdot P\left(\sigma^2\right), \end{aligned} \tag{10}$$

where $K^{(i)} = i \notin (1, \ldots, K)$. The birth-death process is conditional on the prespecified birth rate, λ_b. By setting this birth rate, which controls how often a "birth" of a new component occurs, equal to λ as suggested and done in [1], this computationally allows the death rates to be a likelihood ratio absent of λ. In other words, the likelihood of the data drives the death rates and ultimately the decision of a new cluster. Given that decision is a "birth," the new cluster's parameters $\boldsymbol{\mu_{K+1}}$,σ_{K+1}^2, and π_{K+1} are sampled from their prior distributions:

$$\begin{aligned} \boldsymbol{\mu_{K+1}} &\sim MVN_3\left(\boldsymbol{\xi}, \kappa^{-1}\right) \\ \sigma_{K+1}^2 &\sim IG(20, 0.5) \\ \pi_{K+1} &\sim K(1-\boldsymbol{\pi})^{K-1}. \end{aligned} \tag{11}$$

The mixing proportions are adjusted by multiplying all current proportions by $(1-\pi_{K+1})$ if a birth occurs or dividing by $(1-\pi_i)$ if a death occurs.

To simulate values for all remaining parameters and hyperparameters, we implement the Gibbs sampler. Conditional posterior distributions are listed below. Note that "·" denotes data and other parameters not listed. The conditional posterior of Z_{rj} is

$$\begin{aligned} & P\left(Z_{rj} = k \mid \boldsymbol{\pi}, \theta_{rj}, G_k, \cdot\right) \\ & \quad \propto P\left(\theta_{rj} \mid Z_{rj} = k, G_k, \cdot\right) P\left(G_k \mid \cdot\right) P\left(Z_{rj} = k\right) \end{aligned}$$

$$= \left\{ \frac{\alpha MVN_3 \left(\boldsymbol{\mu_k}, \Sigma_k\right)}{\alpha + n_k - 1} + \frac{\sum_{c=1}^{C_k} \delta_c \left(\theta_{rj}\right)}{\alpha + n_k - 1} \right\} \pi_k, \tag{12}$$

where $c = 1, \ldots, C_k$ is the number of subclusters for cluster $k \in (1, \ldots, K)$, n_k is the number of foci in cluster k, and $\delta_c(\theta_{rj})$ denotes the unit point mass,

$$\boldsymbol{\pi} \mid \mathbf{Z} \sim Dir\left(n_1 + 1, \ldots, n_K + 1\right), \tag{13}$$

and

$$\begin{aligned} P\left(\theta_{rj} \mid k, \cdot\right) &\propto \prod_{Z_{rj} \in k} P\left(s_{rj} \mid p_r, \theta_{rj}, \cdot\right) \\ &\quad \cdot P\left(\theta_{rj} \mid Z_{rj} = k, G_k\right) P\left(G_k \mid G_{0k}, \alpha\right) \\ &\quad \cdot G_{0k}\left(\theta_{rj}; \boldsymbol{\mu_k}, \sigma_k^2\right) = \prod_{Z_{rj} \in k} \left\{MVN_3\left(p_r + \theta_{rj}, \Sigma\right)\right\} \\ &\quad \cdot \left\{ \frac{\alpha MVN_3 \left(\boldsymbol{\mu_k}, \Sigma_k\right)}{\alpha + n_k - 1} + \frac{\sum_{q=1, q \neq rj}^{n_k} \delta_{\theta_q} \left(\theta_{rj}\right)}{\alpha + n_k - 1} \right\}, \end{aligned} \tag{14}$$

which is the distribution of a DP with $\delta_{\theta_q}(\theta_{rj})$ being the unit point mass and n_k the number of foci in some cluster $k \in (1, \ldots, K)$,

$$\begin{aligned} \boldsymbol{\mu_k} \mid \cdot &\sim MVN_3 \left(\left(\kappa + n_k \sigma_k^{-2} I_3\right)^{-1} \right. \\ &\quad \left. \cdot \left(\kappa \boldsymbol{\xi} + n_k \sigma_k^{-2} I_3 \overline{\theta}_{rj}\right), \left(\kappa + n_k \sigma_k^{-2} I_3\right)^{-1}\right), \end{aligned} \tag{15}$$

where $\overline{\theta}_{rj} = \sum_{rj \in k} \theta_{rj} / n_k$ and denotes the average of all θ_{rj} in cluster k,

$$\begin{aligned} \sigma_k^2 \mid \cdot &\sim IG\left(\frac{3 \times n_k}{2} + 20, 0.5 \right. \\ &\quad \left. + \frac{1}{2} \sum_{Z_{rj} \in k} \left(\theta_{rj} - \boldsymbol{\mu_k}\right)' \left(\theta_{rj} - \boldsymbol{\mu_k}\right)\right), \end{aligned} \tag{16}$$

where n_k is the number of foci in cluster k, $k = 1, \ldots, K$. The conditional posterior of p_r is generated from a DP again:

$$\begin{aligned} P\left(p_r \mid \cdot\right) &\propto \prod_{r=1}^{R} P\left(s_{rj} \mid p_r, \theta_{rj}, \cdot\right) P\left(\theta_{rj} \mid \cdot\right) P\left(p_r \mid G_p\right) \\ &\quad \cdot P\left(G_p \mid G_{0p}, \alpha_p\right) G_{0p}\left(p_r; \mu_p, \sigma_p^2\right) \\ &= \prod_{r=1}^{R} \left\{MVN_3\left(p_r + \theta_{rj}, \Sigma\right)\right\} \\ &\quad \cdot \left\{ \frac{\alpha_p TMVN_3\left(\boldsymbol{\mu_p}, \Sigma_p, -l, l\right)}{\alpha_p + R - 1} \right. \\ &\quad \left. + \frac{\sum_{q=1, q \neq r}^{R} \delta_{p_q}\left(p_r\right)}{\alpha_p + R - 1} \right\}, \end{aligned} \tag{17}$$

where $\delta_{p_q}(p_r)$ is the unit point mass. The conditional posterior distributions for the related hyperparameters are

$$\begin{aligned} \boldsymbol{\mu_p} \mid \cdot &\sim MVN_3 \left(\left(1 + R\sigma_p^{-2}\right)^{-1} \right. \\ &\quad \left. \cdot \left(R\sigma_p^{-2}\left(\overline{p_r}\right)\right), \left(1 + R\sigma_p^{-2}\right)^{-1} I_3\right), \end{aligned} \tag{18}$$

where $\overline{p_r} = \sum_{r \in R} p_r / R$ and represents the average group effect and

$$\sigma_p^2 \mid \cdot \sim IG\left(\frac{3 \times R}{2} + 5, \frac{\sum_{r=1}^{R} \left(p_r - \boldsymbol{\mu_p}\right)^2 + 1}{2}\right). \tag{19}$$

Lastly, the sampling distribution for Σ is

$$\begin{aligned} \sigma^2 \mid \cdot &\sim IG\left(\frac{3 \times n + 1}{2}, \right. \\ &\quad \left. \frac{1 + \sum_{Z_{rj} \in n} \left(s_{rj} - p_r - \theta_{rj}\right)' \left(s_{rj} - p_r - \theta_{rj}\right)}{2}\right). \end{aligned} \tag{20}$$

The sampling of unique values for θ_{rj} and p_r can be performed using Neal's algorithm 8 [14]. It works by introducing m auxiliary parameters that are independent to other parameters to represent potential values for θ_{rj} and p_r [14]. Algorithm 8 for updating clustering assignments, denoted as c, is as follows:

(i) The state of the Markov chain consists of $c = \{c_1, \ldots, c_n\}$ and $\Phi = (\phi_c; c \in c_1, \ldots, c_n)$ with ϕ_c denoting cluster parameters, e.g., θ_c in our application. Repeatedly sample as follows:

(ii) For $i = 1, \ldots, n$, let k^- be the number of distinct c_l for $l \neq i$, and $h = k^- + m$. Label these c_l with values in $\{1, \ldots, k^-\}$. If $c_i = c_l$ for some $l \neq i$, draw values independently from base distribution G_0 for those ϕ_c for which $k^- < c \leq h$. If $c_i \neq c_l$ for all $l \neq i$, let c_i have the label $k^- + 1$, and draw values independently from G_0 for those ϕ_c for which $k^- + 1 < c \leq h$. Draw a new value for c_i from $\{1, \ldots, h\}$ using the following probabilities:

$$\begin{aligned} &P\left(c_i = c \mid c_{-i}, y_i, \phi_1, \ldots, \phi_h\right) \\ &\quad = \begin{cases} \dfrac{n_{-i,c}}{n - 1 + \alpha} F\left(y_i, \phi_c\right) & \text{for } 1 \leq c \leq k^- \\ \dfrac{(\alpha/m)}{n - 1 + \alpha} F\left(y_i, \phi_c\right) & \text{for } k^- < c \leq h, \end{cases} \end{aligned} \tag{21}$$

where $F(y_i, \theta_c)$ is the likelihood with θ_c and observation i, y_i, involved. In our case, it is s_{ij},

(iii) where $n_{-i,c}$ is the number of c_l for $l \neq i$ that are equal to c and b is the appropriate normalizing constant. Change the state to contain only those ϕ_c that are now associated with one or more observation.

(iv) For all $c \in \{c_1, \ldots, c_n\}$, draw new values from $\phi_c \mid y_i$ such that $c_i = c$, or perform some other update to ϕ_c that leaves this distribution invariant [14].

Thus, the value for θ_{rj} for those foci in cluster k and subcluster c may be sampled from

$$\theta_{k,c} \mid \cdot \sim MVN_3 \left(\left(\sigma_k^{-2} + n_{k,c}\sigma^{-2} \right)^{-1} \cdot \left(\sigma_k^{-2} \boldsymbol{\mu}_{\boldsymbol{k}} + n_{k,c}\sigma^{-2} \overline{\left(s_{rj} - p_r \right)} \right), \left(\sigma_k^{-2} + n_k \sigma^{-2} \right)^{-1} \cdot I_3 \right), \quad (22)$$

where $n_{k,c}$ are the number of foci in some cluster k and subcluster c and $\overline{(s_{rj} - p_r)}$ denotes the mean of the observed data after adjusting for group effect. The value for p_r for those groups (studies) in group cluster c (different c notation than subclusters above) may be sampled from

$$p_c \mid \cdot \sim TMVN_3 \left(\left(\sigma_p^{-2} + n_c\sigma^{-2} \right)^{-1} \cdot \left(\sigma_p^{-2} \boldsymbol{\mu}_{\boldsymbol{p}} + n_c\sigma^{-2} \overline{\left(s_{rj} - \theta_{rj} \right)} \right), \left(\sigma_p^{-2} + n_c \sigma^{-2} \right)^{-1} I_3, -l, l \right), \quad (23)$$

where n_c is the number of groups in cluster c and similarly $\overline{(s_{rj} - \theta_{rj})}$ denotes the mean of the observed data after adjusting for individual effect.

2.4. Determining the Clusters. To estimate the number of clusters and the center of each cluster and cluster assignment, we implement the same least-squared Euclidean distance method introduced in [15] and used previously in our work in [10] and reiterated below. This method draws inferences on clusters based on a set of converged MCMC iterations and chooses one iteration as the final estimates for the clusters and related parameters. This final MCMC iteration is selected due to its smallest Euclidean distance to the expected cluster assignments estimated based on a set of independent converged MCMC iterations. This approach incorporates all clustering information in the MCMC sampling process [15]:

(1) After the prespecified number of MCMC burn-ins, let the MCMC simulations continue for an additional W iterations. An averaged clustering matrix is then created, denoted as A, and is an $n \times n$ matrix with each block or $(i, j)^{th}$ entry denoting how often foci i and j $(i, j = 1, \ldots, n)$ are in the same cluster. Specifically, each $(i, j)^{th}$ entry is the proportion of the W iterations that two foci are in the same cluster.

(2) Let the MCMC run additional F iterations, where, for each iteration,

   (a) create an $n \times n$ matrix using indicators to denote which foci are clustered together; i.e., let the $(i, j)^{th}$ entry denote a 1 if foci i and j are in one cluster and 0 otherwise.

   (b) use Euclidean distance to determine the similarity between this indicator matrix and the averaged clustering matrix A.

(3) Among the F iterations, select the iteration and respective clustering pattern, number of clusters, and parameters that produce the smallest Euclidean distance.

3. Inference

3.1. Simulation Settings. Simulations were utilized to illustrate and assess the proposed method. We assumed an fMRI metadata study setting including 50 studies, with each study containing 10 foci. Collectively these foci were simulated from three clusters centered as (x,y,z) talairach coordinates at $(1, 1, 1)^T$, $(2, 2, 2)^T$, and $(4, 4, 4)^T$ containing 150, 150, and 200 foci, respectively. It was also assumed that half the data, 250 foci or half from each individual cluster, came from one study cluster centered at $(0.1, 0.1, 0.1)^T$ and the remaining data from a second study cluster centered at $(0.4, 0.4, 0.4)^T$. These study clusters were linear shifts to cluster centers. For example, 75 foci in cluster one were centered at $(1.1, 1.1, 1.1)^T$ and the other half were centered at $(1.4, 1.4, 1.4)^T$; the study effect was a linear shift to all three dimensions from the cluster center. We made various alterations to this general setting:

(1) Normal setting: We simulated data from a multivariate normal for each cluster with respective means described above and a variance of $\Sigma = 0.002 I_3$. This creates spheres with little variation and we expect the method to have the ability to correctly identify the clusters.

(2) Chi-squared (skewed) setting: The method's ability to cluster in the presence of abnormal patterns is an important factor in spatial clustering. For this setting, we applied the same scenario as in the normal setting in respect of clusters 1 and 2 but simulated cluster 3 using a chi-squared distribution with 4 degrees of freedom.

(3) Large variance setting: The last scenario is designed to assess the robustness of the method with respect to the distance between and among clusters. To this end, we applied the normal setting but considered increasing levels (referred to large1, large2, large3, and large4 settings, resp.) of Σ: $\Sigma = 0.01 I_3, 0.05 I_3, 0.1 I_3$, and $0.2 I_3$, representing gradually closer distances among clusters.

For each setting, we implemented a grid search for a single dataset to estimate values of α_p and α based on the minimization of DIC. We let precision parameter values be 0.01, 0.05, 0.1, 0.5, 1, 2, and 5. Based on α_p and α estimates, 100 MC datasets were generated with 2500 burn-in iterations, 500 working iterations to calculate the probability matrix for determining the clusters, and 100 additional iterations

Table 1: Simulation assessments for three foci-level clusters.

Scenario (α_p, α)	Median Num. of Clusters (SD)∗	Cluster index	Average Sensitivity (SD)∗	Average Specificity (SD)∗	Average Accuracy (%) (SD)∗
Normal (0.1, 1.0)	∗∗IC: 3 (0.88)	1	0.96 (0.13)	1 (0)	0.96 (0.13)
		2	0.96 (0.13)	1 (0)	
		3	0.96 (0.13)	1 (0)	
	∗∗SC: 2 (0.32)	1	0.93 (0.26)	1 (0.04)	0.96 (0.13)
		2	1 (0.04)	0.93 (0.26)	
Chi-squared (0.05, 0.1)	IC: 19 (2.17)	1	0.94 (0.09)	0.91 (0.15)	0.54 (0.13)
		2	0.74 (0.43)	0.96 (0.02)	
		3	0.09 (0.03)	0.94 (0.16)	
	SC: 2 (0.56)	1	0.76 (0.37)	0.9 (0.1)	0.82 (0.18)
		2	0.88 (0.11)	0.78 (0.36)	
Large 1 (0.05, 0.1)	IC: 3 (0.82)	1	0.96 (0.13)	1 (0)	0.96 (0.13)
		2	0.96 (0.13)	1 (0)	
		3	0.96 (0.13)	1 (0)	
	SC: 2 (0.29)	1	0.92 (0.27)	1 (0)	0.96 (0.14)
		2	1 (0)	0.92 (0.27)	
Large 2 (0.01, 0.5)	IC: 3 (0.77)	1	0.99 (0.04)	1 (0)	0.99 (0.02)
		2	0.99 (0.03)	1 (0)	
		3	1 (0.01)	1 (0)	
	SC: 2 (0.14)	1	1 (0.02)	1 (0.01)	1 (0.01)
		2	1 (0.01)	1 (0)	
Large 3 (0.05, 0.05)	IC: 5 (1.25)	1	0.91 (0.13)	1 (0)	0.92 (0.07)
		2	0.88 (0.16)	1 (0)	
		3	0.96 (0.09)	1 (0)	
	SC: 2 (0.37)	1	0.97 (0.06)	0.98 (0.03)	0.97 (0.04)
		2	0.98 (0.03)	0.98 (0.03)	
Large 4 (0.01, 2.0)	IC: 9 (1.3)	1	0.66 (0.14)	0.99 (0.01)	0.62 (0.09)
		2	0.33 (0.08)	0.99 (0.01)	
		3	0.8 (0.17)	1 (0)	
	SC: 1 (0.36)	1	0.04 (0.17)	0.99 (0.05)	0.51 (0.08)
		2	0.98 (0.07)	0.05 (0.18)	

∗SD: standard deviation across 100 MC replicates; ∗∗IC: individual foci cluster; SC: study effect clusters.

to infer the number of clusters and individual foci cluster centers.

Model assessment consists of three evaluations: sensitivity, specificity, and accuracy. Sensitivity and specificity are defined by their generic definitions, the proportion of foci that are correctly assigned to their simulated cluster, and the proportion of foci that are correctly not assigned their nonsimulated cluster. Accuracy is defined as the percentage of foci that are correctly clustered. Note that the definition of accuracy takes into account both true positive and true negatives. In addition to our methodology, we applied a very common existing clustering approach for continuous data, K-means, to our simulation settings. Although this method cannot adjust for additional covariates, it allows for a comparison to existing methods. Lastly, to highlight the advantage of using a mixture of DPs over existing clustering approaches, we applied our approach, a revised version of our approach using a mixture of multivariate normal distributions rather than DPs, and Kmeans to the normal and chi-squared simulation scenarios. As the emphasis was on clustering performance, the group effect was assumed to be known for these two settings.

3.2. Simulation Results. Table 1 summarizes the findings on the three foci-level cluster identifications and the quality of the identified clusters. The proposed method gives high sensitivities and specificities across all scenarios. The accuracy of cluster assignment overall is higher than 90% when the variation in the data is relatively small and only dropping once clusters were large enough to overlap (scenario Large 4). The proposed methodology was also accurate at identifying the correct number of clusters as indicated by the median

Table 2: Kmeans simulation assessments for three foci-level clusters.

Scenario:	Median Num. of Clusters (SD)∗	Cluster index	Average Sensitivity (SD)∗	Average Specificity (SD)∗	Average Accuracy (%) (SD)∗
Normal	5 (0.95)	1	0.73 (0.29)	0.98 (0.09)	0.7 (0.12)
		2	0.72 (0.3)	0.98 (0.08)	
		3	0.67 (0.26)	1 (0)	
Chi-squared	10 (0.5)	1	0.58 (0.19)	0.99 (0.01)	0.5 (0.09)
		2	0.86 (0.28)	0.92 (0.03)	
		3	0.16 (0.08)	0.98 (0.11)	
Large 1	5 (0.98)	1	0.68 (0.33)	1 (0.04)	0.7 (0.13)
		2	0.76 (0.26)	0.96 (0.12)	
		3	0.67 (0.26)	1 (0)	
Large 2	7 (1.29)	1	0.61 (0.27)	0.99 (0.06)	0.58 (0.11)
		2	0.61 (0.27)	0.99 (0.06)	
		3	0.54 (0.27)	1 (0)	
Large 3	9 (0.94)	1	0.49 (0.19)	1 (0)	0.46 (0.08)
		2	0.47 (0.17)	1 (0)	
		3	0.43 (0.24)	1 (0)	
Large 4	10 (0.57)	1	0.43 (0.16)	1 (0.01)	0.41 (0.07)
		2	0.42 (0.14)	1 (0.01)	
		3	0.38 (0.2)	1 (0)	

∗SD: standard deviation across 100 MC replicates.

number of individual and study clusters. In comparison, results from the Kmeans approach indicate relatively lower statistics in sensitivity and specificity (Table 2). The accuracy is around 70%, when the variations in the data are relatively small. However, compared to the proposed method, the Kmeans approach often inferred a higher number of clusters as indicated by the larger median number of clusters. The computation time of a single dataset for the mixture of DPs took, on average, 7-8 hours on a high performance computer (Dell cluster with 88 compute nodes, 3120 total central processing unit cores, 20664 Giga-bytes of RAM, and 61440 total graphic processing unit cores).

After removing study effects, the comparison between the mixture of DPs, mixture of normal, and Kmeans is as expected (Table 3). Both mixtures performed exceptionally well at identifying the three normally distributed clusters while the Kmeans performance was adequate with an overall accuracy of 80% (compared to 100% for both mixtures). Once the data deviated from normality, the mixture of normals approach was unable to differentiate the clusters resulting in low accuracy (32%). The Kmeans approach performed similarly to the mixture of DPs with both approaches having low sensitivities for the third cluster which was skewed but with the mixture of DPs resulting in superior accuracy. The mixture of DPs was able to differentiate clusters 1 and 2 as indicated by 99% sensitivity and 100% specificity measures but tended to “overcluster” cluster 3 into smaller clusters as indicated by the large median number of clusters, 14% sensitivity, and 100% specificity. In regard to overall accuracy, the mixture of DPs outperformed the mixture of normals and Kmeans when the data are skewed.

4. Real Data Application

For this application, we applied the proposed method to a meta-analysis dataset. Constructed originally in [16], this data consists of a total of 162 neuroimaging publications, of which 57 were PET and 105 were fMRI. Among these 162 publications, there were 437 contrasts or studies. Only those foci that were deemed significantly activated by their study specific criteria were included for a total of 2,478 foci. Summary statistics for this data can be seen in Tables 4 and 5.

As with the simulation studies, grid search and DIC were used to estimate values for α_p and α. Potential precision parameters values were 0.01, 0.05, 0.1, 0.5, 1, 2, 5, and 7.5. Each combination was performed over 2,600 iterations, 2,000 of those for burn-in, 500 for the probability matrix calculation, and final 100 to infer individual clusters and their centers. To assist with the magnitude of the likelihood calculations, the data was scaled down by 10.

It was found that the precision parameter combination of $\alpha_p = 0.05$ and $\alpha = 1$ produced the smallest DIC. Convergence, with the initial 2,000 discarded, was checked based on trace plots. Based on the proposed method, we identified four study clusters and 14 individual foci clusters. A single DIC setting for this data took, on average, 72 hours to run on the HPC. The break down of the 14

Table 3: Comparison of approaches for selected simulated settings assuming study effect is known.

Setting	Approach	Median Num. of Clusters (SD)∗	Cluster index	Average Sensitivity (SD)∗	Average Specificity (SD)∗	Average Accuracy (%) (SD)∗
Normal	DP	3 (0)	1	1 (0)	1 (0)	1 (0)
			2	1 (0)	1 (0)	
			3	1 (0)	1 (0)	
	Mixture	3 (0)	1	1 (0)	1 (0)	1 (0)
			2	1 (0)	1 (0)	
			3	1 (0)	1 (0)	
	Kmeans	3 (0.6)	1	0.81 (0.36)	0.93 (0.16)	0.8 (0.22)
			2	0.79 (0.38)	0.94 (0.15)	
			3	0.79 (0.25)	1 (0)	
Chi-squared	DP	16 (1.76)	1	0.99 (0.01)	0.99 (0.01)	0.65 (0.01)
			2	0.99 (0.01)	0.97 (0.01)	
			3	0.14 (0.03)	1 (0)	
	Mixture	1 (0.29)	1	0.06 (0.24)	0.98 (0.11)	0.32 (0.1)
			2	0.95 (0.22)	0.06 (0.24)	
			3	0.05 (0.21)	1 (0)	
	Kmeans	10 (0.22)	1	0.99 (0.06)	0.98 (0.01)	0.62 (0.1)
			2	0.87 (0.34)	0.93 (0.03)	
			3	0.16 (0.06)	0.94 (0.17)	

∗SD: standard deviation across 100 MC replicates.

Table 4: Descriptive statistics∗.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Number of foci per pub.	1.00	5.75	10.00	15.11	17.25	110.00
Number of foci per study	1.00	2.00	4.00	5.67	7.00	47.00
Number of subjects per pub.	4.00	9.00	11.00	12.26	14.00	40.00
Number of studies per pub.	1.000	1.000	2.000	2.67	4.000	12.000

∗Min: minimum, 1st Qu: 25% percentile, 3rd Qu: 75% percentile, Max: maximum, pub: publication.

Table 5: Frequency of emotions.

Emotions	Frequency of studies (% of total studies)	Frequency of foci (% of total foci)
aff∗	175 (40.05%)	881 (35.55%)
anger	26 (5.95 %)	166 (6.7%)
disgust	44 (10.07%)	337 (13.6%)
fear	68 (15.56%)	367 (14.81%)
happiness	36 (8.24%)	178 (7.18%)
mixed	41 (9.38%)	195 (7.87%)
sadness	45 (10.3%)	348 (14.04%)
surprise	2 (0.46%)	6 (0.24%)
Total	437	2478

∗aff: affective.

individual foci clusters by center location, brain location, foci frequency, and study frequency can be seen in Table 6. The frequency of each foci-associated emotion within each of the 14 clusters can be seen in Table 7. The affective emotion was dominating in all clusters with fear being the second dominating emotion in clusters 1, 2, 3, 11, and 13, disgust in clusters 5 and 14, sadness in clusters 6 and 10, and a mixture of emotions in the remaining clusters. When only focusing on those foci that fell within known brain regions of interest, as seen in Table 8, the dominating emotion, other

Table 6: Meta-data cluster results.

Cluster Centers	Brian Regions	Cluster Index	# of foci per cluster (% of total foci)	# of studies per cluster (% of all studies)
(-9.11, -14.45, 1.4)	Temporal Mid L	1	1615 (65.17)	386 (88.33)
(-0.83, -4.64, 0.54)	Temporal Mid L	2	188 (7.59)	139 (31.81)
(16.75, -10.63, -11.33)	Temporal Mid R	3	148 (5.97)	117 (26.77)
(3.48, -5.52, -2.15)	Temporal Inf L	4	102 (4.12)	84 (19.22)
(2.08, -4.61, -0.14)	Temporal Mid R	5	77 (3.11)	67 (15.33)
(0.87, -7.7, -0.63)	Temporal Pole Sup L	6	67 (2.7)	59 (13.5)
(0.99, -6.33, 2.62)	Cerebelum 6 R	7	55 (2.22)	43 (9.84)
(2.3, -5.89, 0.32)	NA	8	54 (2.18)	46 (10.53)
(0.43, -6.59, 0.6)	Postcentral L	9	41 (1.65)	36 (8.24)
(-0.18, -4.99, 0.11)	Cerebelum 6 R	10	39 (1.57)	38 (8.7)
(1.69, -5.25, 1.43)	Temporal Sup R	11	38 (1.53)	34 (7.78)
(-0.64, -5.93, 1.77)	Postcentral L	12	22 (0.89)	20 (4.58)
(1.29, -5.28, 1.23)	Precentral R	13	18 (0.73)	17 (3.89)
(0.45, -5.31, -0.87)	Occipital Inf R	14	14 (0.56)	14 (3.2)

R: right hemisphere, L: left hemisphere.

than affective, was sadness, fear, and disgust, respectively. When compared to the number of clusters identified by the spatial Cox Point process (53) and Kmeans (20) performed in [10], fewer clusters were identified with our current application. It should be noted that this particular data does not visually indicate distinct clusters and is closer to a more uniform distribution throughout the brain which may lead to an inaccurate number of identifiable clusters. However, given the results from our previous analysis in [10] and findings in simulation studies, it is possible that the clusters formed were rather subtle and actually might not be distinct enough.

5. Conclusion and Discussion

Modeling the realization of observed foci as a linear association of study effect and individual foci cluster effect with a multivariate normal random error was motivated by the limitation of the spatial Cox process to statistically distinguish between a cluster and a mode or peak of a cluster. The overall aim remained to identify activated regions within the brain using fMRI coordinate-based metadata. By modeling the data in this fashion, it was hopeful that the distribution could statistically differentiate between clusters and modes of clusters while retaining the flexibility and robustness to mimick the behavior of the data.

Simulation studies demonstrated that the method can fit data generated from normal or abnormal distributions. Furthermore, it was able to identify clusters within covariates while retaining the integrity to identify individual clusters. Both the proposed method and Kmeans were unable to correctly identify clusters when they were large and overlapped and both the mixture of normals and mixture of DPs performed poorly at identifying a cluster severely skewed.

When applied to a fMRI metadataset, the method identified a relatively low number of clusters. Given the low sensitivity findings in the simulated study with high noise, it can be concluded that this data had a high likelihood of being too broad. When the same data was analyzed with the spatial Cox process, the difference in the results was extreme. Not only were the number of clusters substantially less, but also none of the cluster centers identified from the proposed method came close to those identified in the first method. It is worth mentioning that the meta-analysis data is not distinctly grouped and is more uniformly distributed throughout the brain and perhaps the model used did not provide the best fit.

The primary advantage to this method, besides its flexibility, is its ability to describe irregular spatial patterns and its sampling design to statistically differentiate clusters. Because of its adaptable nature, this model can also adjust for any covariate(s) of interest. However, based on simulation studies and the fMRI metadata application, the proposed method tends to be too insensitive and has a difficult time identifying clusters when data are not distinctly differentiated. A potential limitation in the approach is that each DP within the mixture was assumed to have the same precision parameter. It was noted that, during the simulation studies when the mixture of DPs was attempting to fit the Chi-squared simulations (without study effect), it was overclustering the skewed cluster. However, when the precision parameter was smaller, the identifiability of cluster 3 became more accurate but became more inaccurate for clusters 1 and 2. Thus, to further improve flexibility and accuracy for this approach when data is skewed, each DP potentially requires its own unique precision parameter. Furthermore, this method's clustering ability is limited by the identification of study effects which may be improved by implementing stronger restrictions or could be an effect of having multiple DPs. Our future work will focus on these issues, allowing study effect to be random rather than a fixed effect, and identifying if a large number of DPs within the model is indeed a limitation.

Table 7: Breakdown of emotions and their frequencies by individual foci cluster∗.

Cluster Index: Total foci in that cluster					
Emotion Frequency of emotion (% of total cluster foci)					
Cluster: 1	1615	**Cluster: 2**	188	**Cluster: 3**	148
aff	562 (34.8)	aff	57 (30.32)	aff	53 (35.81)
anger	110 (6.81)	anger	10 (5.32)	anger	12 (8.11)
disgust	225 (13.93)	disgust	18 (9.57)	disgust	22 (14.86)
fear	234 (14.49)	fear	35 (18.62)	fear	31 (20.95)
happiness	121 (7.49)	happiness	17 (9.04)	happiness	8 (5.41)
mixed	131 (8.11)	Mixed	18 (9.57)	mixed	8 (5.41)
sadness	228 (14.12)	sadness	33 (17.55)	sadness	13 (8.78)
surprise	4 (0.25)			surprise	1 (0.68)
Cluster: 4	102	**Cluster: 5**	77	**Cluster: 6**	67
aff	44 (43.14)	aff	35 (45.45)	aff	22 (32.84)
anger	6 (5.88)	anger	3 (3.9)	anger	4 (5.97)
disgust	15 (14.71)	disgust	13 (16.88)	disgust	5 (7.46)
fear	9 (8.82)	fear	5 (6.49)	fear	14 (20.9)
happiness	9 (8.82)	happiness	5 (6.49)	happiness	6 (8.96)
mixed	4 (3.92)	Mixed	7 (9.09)	Mixed	4 (5.97)
sadness	15 (14.71)	sadness	9 (11.69)	sadness	11 (16.42)
				surprise	1 (1.49)
Cluster: 7	55	**Cluster: 8**	54	**Cluster: 9**	41
aff	21 (38.18)	aff	19 (35.19)	aff	13 (31.71)
anger	7 (12.73)	anger	2 (3.7)	anger	5 (12.2)
disgust	7 (12.73)	disgust	9 (16.67)	disgust	8 (19.51)
fear	6 (10.91)	fear	7 (12.96)	fear	8 (19.51)
happiness	3 (5.45)	mixed	9 (16.67)	happiness	2 (4.88)
mixed	4 (7.27)	sad	8 (14.81)	mixed	1 (2.44)
sadness	7 (12.73)			sad	4 (9.76)
Cluster: 10	39	**Cluster: 11**	38	**Cluster: 12**	22
aff	15 (38.46)	aff	20 (52.63)	aff	9 (40.91)
anger	2 (5.13)	anger	1 (2.63)	anger	2 (9.09)
disgust	5 (12.82)	disgust	2 (5.26)	disgust	3 (13.64)
fear	3 (7.69)	fear	7 (18.42)	fear	3 (13.64)
happiness	3 (7.69)	happiness	2 (5.26)	mixed	2 (9.09)
mixed	4 (10.26)	sad	6 (15.79)	sad	3 (13.64)
sadness	7 (17.95)				
Cluster: 13	18	**Cluster: 14**	14		
aff	5 (27.78)	aff	6 (42.86)		
anger	2 (11.11)	disgust	3 (21.43)		
disgust	2 (11.11)	happiness	1 (7.14)		
fear	5 (27.78)	mixed	2 (14.29)		
happiness	1 (5.56)	sad	2 (14.29)		
mixed	1 (5.56)				
sadness	2 (11.11)				

Table 8: Breakdown of emotions and their frequencies by individual foci cluster for ROI∗.

Cluster Index: Total foci in that cluster Emotion Frequency of emotion (% of total cluster foci)					
Cluster: 1	489	**Cluster: 2**	63	**Cluster: 3**	54
aff	217 (44.38)	aff	29 (46.03)	aff	21 (38.89)
anger	20 (4.09)	anger	3 (4.76)	anger	3 (5.56)
disgust	68 (13.91)	disgust	8 (12.7)	disgust	4 (7.41)
fear	63 (12.88)	fear	13 (20.63)	fear	8 (14.81)
happiness	28 (5.73)	happiness	2 (3.17)	happiness	6 (11.11)
mixed	22 (4.5)	mixed	4 (6.35)	mixed	5 (9.26)
sadness	71 (14.52)	sadness	4 (6.35)	sadness	7 (12.96)
Cluster: 4	34	**Cluster: 5**	25	**Cluster: 6**	20
aff	15 (44.12)	aff	13 (52)	aff	8 (40)
anger	1 (2.94)	anger	1 (4)	disgust	1 (5)
disgust	4 (11.76)	disgust	4 (16)	fear	4 (20)
fear	3 (8.82)	fear	1 (4)	happiness	1 (5)
happiness	1 (2.94)	happiness	2 (8)	mixed	1 (5)
mixed	1 (2.94)	mixed	2 (8)	sadness	5 (25)
sadness	9 (26.47)	sadness	2 (8)		
Cluster: 7	19	**Cluster: 8**	17	**Cluster: 9**	16
aff	6 (31.58)	aff	6 (35.29)	aff	9 (56.25)
anger	1 (5.26)	anger	2 (11.76)	fear	3 (18.75)
disgust	4 (21.05)	disgust	3 (17.65)	happiness	2 (12.5)
mixed	2 (10.53)	happiness	1 (5.88)	sadness	2 (12.5)
sadness	6 (31.58)	mixed	1 (5.88)		
		sadness	4 (23.53)		
Cluster: 10	15	**Cluster: 11**	9	**Cluster: 12**	7
aff	6 (40)	aff	3 (33.33)	aff	3 (42.86)
anger	1 (6.67)	anger	1 (11.11)	anger	1 (14.29)
disgust	2 (13.33)	disgust	2 (22.22)	disgust	1 (14.29)
fear	2 (13.33)	fear	1 (11.11)	fear	2 (28.57)
happiness	1 (6.67)	sadness	2 (22.22)		
mixed	2 (13.33)				
sadness	1 (6.67)				
Cluster: 13	6	**Cluster: 14**	5		
aff	3 (50)	aff	3 (60)		
disgust	1 (16.67)	happiness	1 (20)		
sadness	2 (33.33)	sadness	1 (20)		

∗ROI: region of interest; aff: affective.

Data Availability

Data is not currently publicly available but available upon request from Professor Tor D. Wager at the University of Colorado.

Conflicts of Interest

No conflicts of interest exist for any author.

Acknowledgments

Dr. Kang's effort was supported by an NIH Grant 1R01MH105561. Dr. Zhang and Dr. Ray's efforts were supported by their start funds provided by the University of Memphis.

References

[1] M. Stephens, "Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods," *The Annals of Statistics*, vol. 28, no. 1, pp. 40–74, 2000.

[2] M. Aitkin and R. Healey, "Estimation and hypothesis testing in finite mixture models. Journal of the Royal Statistical Society," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 67–75, 1985.

[3] G. J. McLachlan and K. E. Basford, *Mixture Models: Applications to Clustering*, Marcel Dekker, New York, NY, USA, 1988.

[4] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 381–396, 2002.

[5] J.-M. Marin, K. Mengersen, and C. . Robert, "Bayesian modelling and inference on mixtures of distributions," in *Bayesian thinking: modeling and computation*, vol. 25 of *Handbook of Statist.*, pp. 459–507, Elsevier/North-Holland, Amsterdam, 2005.

[6] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

[7] P. J. Green and D. I. Hastie, "Reversible jump mcmc," *Genetics*, vol. 1550, no. 3, pp. 1391–1403, 2009.

[8] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 577–588, 1995.

[9] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, vol. 2, pp. 1152–1174, 1974.

[10] M. Ray, J. Kang, and H. Zhang, "Identifying Activation Centers with Spatial Cox Point Processes Using fMRI Data," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 6, pp. 1130–1141, 2016.

[11] P. Congdon, *Bayesian statistical modelling*, vol. 704, John Wiley & Sons, 2007.

[12] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 4, pp. 583–639, 2002.

[13] C. Preston, "Spatial birth-and-death processes," *Advances in applied probability*, vol. 70, no. 03, pp. 405–408, 1975.

[14] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[15] D. B. Dahl, "Model-based clustering for expression data via a dirichlet process mixture model," *Bayesian Inference for Gene Expression and Proteomics*, pp. 201–218, 2006.

[16] H. Kober, L. F. Barrett, J. Joseph, E. Bliss-Moreau, K. Lindquist, and T. D. Wager, "Functional grouping and cortical-subcortical interactions in emotion: A meta-analysis of neuroimaging studies," *NeuroImage*, vol. 42, no. 2, pp. 998–1031, 2008.

Hindawi
Journal of Probability and Statistics
Volume 2018, Article ID 8654173, 11 pages
https://doi.org/10.1155/2018/8654173

Research Article

A Bayesian Adaptive Design in Cancer Phase I Trials Using Dose Combinations in the Presence of a Baseline Covariate

Márcio Augusto Diniz, Sungjin Kim, and Mourad Tighiouart

Biostatistics and Bioinformatics Research Center, Cedars-Sinai Medical Center 8700 Beverly Blvd, Los Angeles, CA 90048, USA

Correspondence should be addressed to Mourad Tighiouart; mourad.tighiouart@cshs.org

Received 3 May 2018; Revised 27 July 2018; Accepted 29 August 2018; Published 1 November 2018

Academic Editor: Ash Abebe

A Bayesian adaptive design for dose finding of a combination of two drugs in cancer phase I clinical trials that takes into account patients heterogeneity thought to be related to treatment susceptibility is described. The estimation of the maximum tolerated dose (MTD) curve is a function of a baseline covariate using two cytotoxic agents. A logistic model is used to describe the relationship between the doses, baseline covariate, and the probability of dose limiting toxicity (DLT). Trial design proceeds by treating cohorts of two patients simultaneously using escalation with overdose control (EWOC), where at each stage of the trial, the next dose combination corresponds to the α quantile of the current posterior distribution of the MTD of one of two agents at the current dose of the other agent and the next patient's baseline covariate value. The MTD curves are estimated as function of Bayes estimates of the model parameters at the end of trial. Average DLT, pointwise average bias, and percent of dose recommendation at dose combination neighborhoods around the true MTD are compared between the design that uses the covariate and the one that ignores the baseline characteristic. We also examine the performance of the approach under model misspecifications for the true dose-toxicity relationship. The methodology is further illustrated in the case of a prespecified discrete set of dose combinations.

1. Introduction

Despite the promise observed in preclinical experiments and initial high response rates, a large number of targeted drugs have not been successful in providing reproducible improvements in survival in patients with cancer when used as single agents. [1] In addition, targeted therapies do not work for every patient since they rely on the presence of the target. Therefore, chemotherapy and radiotherapy approaches are still the backbone of cancer treatment for tumors after surgical excision. These conventional cancer therapies may be combined with targeted agents to enhance treatment efficacy.

Statistical methodologies for designing phase I clinical trials for drug combinations have been studied extensively in the past decade [2–13]. These methods assume that the patient population is homogeneous of treatment tolerance and every patient should be treated at a dose combination corresponding to a predefined target probability of DLT (dose limiting toxicity). Therefore, an additional layer of complexity in specifying the dose-toxicity relationship given a baseline covariate is needed for drug combinations.

Strategies of drug allocation that accommodate individual patient needs have been used in [14–18] for single agent trials. Statistical designs allowing individualized maximum tolerable dose (MTD) determination in single agent cancer phase I trials have also been proposed and implemented in real trials by a number of authors for two groups with no prior knowledge of ordering [19, 20], for two prior ordered groups [21, 22] and two or more prior partially ordered groups [23, 24]. In general, ignoring the heterogeneity between groups can lead to higher toxicities in the most severely impaired group, statistical bias, and inefficiency of the MTD estimate for both groups.

In this work, we extend the design described by Tighiouart et al. [25] using escalation with overdose control (EWOC) principle [26], by treating cohorts of two patients simultaneously and accounting for patient baseline binary covariate. We assume that we do not have prior knowledge of the ordering between groups, but they will be ordered in the sense that the probability of toxicity for one group is always a constant shift from the probability of toxicity for the second group at the same dose. In this way, patients with different covariate values will have parallel MTD curves. This assumption is mathematically convenient and allows us to use parsimonious models due to the small sample size constraints in cancer phase I trials.

This paper is organized as follows. Section 2 will describe the dose-toxicity model and trial design for continuous dose levels. In Section 3, we evaluate the performance of the proposed method by assessing the safety of the trial design and the efficiency of the estimate of the MTD curve. The methodology is extended for discrete dose combinations in Section 4. Discussions will be presented in Section 5.

2. Model

2.1. Dose-Toxicity Model. We propose a parametric model to identify tolerable dose combinations of two synergistic drugs A and B [10–12, 25, 27] given a patient with a binary baseline covariate value of z:

$$\operatorname{Prob}(\delta = 1 \mid x, y, z) = F(\mu + \beta x + \gamma y + \lambda z + \eta x y), \quad (1)$$

where δ is the indicator of DLT, (x, y) are the continuous dose levels of agents A and B, respectively, assuming values in $[X_{min}, X_{max}] \times [Y_{min}, Y_{max}]$, z is a binary baseline covariate value, and F is a known cumulative distribution function.

We assume partial ordering of the probability of DLT, i.e., it is a nondecreasing function of the dose of any one of the agents when the other one is held constant for $z = 0, 1$ and we also assume synergism between the two drugs. These assumptions are translated into constrains in the parameter space given by $\beta, \gamma > 0$, and $\eta \geq 0$, respectively. The MTD C_z for a patient with covariate value z is defined as the set of combinations (x^*, y^*) such that

$$\operatorname{Prob}(\delta = 1 \mid x^*, y^*, z) = \theta. \quad (2)$$

The target probability of DLT, θ, is set relatively high when the DLT is a reversible or nonfatal condition, and low when it is life threatening. Using (1) and (2), the MTD C_z is

$$C_z = \left\{ (x^*, y^*) \in [0,1]^2 : y^* = \frac{F^{-1}(\theta) - \mu - \beta x^* - \lambda z}{\gamma + \eta x^*} \right\}. \quad (3)$$

We reparametrize model (1) to allow a more meaningful prior elicitation. Assuming that $[X_{min}, X_{max}] \times [Y_{min}, Y_{max}]$ will be standardized to be in $[0, 1] \times [0, 1]$, ρ_{000}, the probability of DLT at the minimum available doses of agents A and B for a patient with covariate value $z = 0$; ρ_{100}, the probability of DLT when the level of drug A is X_{max}, the level of drug B is Y_{min} and $z = 0$; ρ_{101}, the probability of DLT when the level of drug A is X_{max}, the level of drug B is Y_{min} and $z = 1$; ρ_{010}, the probability of DLT when the level of drug A is X_{min}, the level of drug B is Y_{max} and $z = 0$; and the interaction parameter η. It follows that

$$\begin{aligned}
\mu &= F^{-1}(\rho_{000}) \\
\beta &= F^{-1}(\rho_{100}) - F^{-1}(\rho_{000}) \\
\gamma &= F^{-1}(\rho_{010}) - F^{-1}(\rho_{000}) \\
\lambda &= F^{-1}(\rho_{101}) - F^{-1}(\rho_{100}).
\end{aligned} \quad (4)$$

Notice that $\beta, \gamma > 0$ implies that $\rho_{000} < \min(\rho_{100}, \rho_{010})$. The MTD set given in (3) can be presented as

$$C_z = \left\{ (x^*, y^*) \in [0,1]^2 : y^* = \frac{G(\theta, \rho_{000}) - (G(\rho_{100}, \rho_{000}))\, x^* - (G(\rho_{101}, \rho_{100}))\, z}{G(\rho_{010}, \rho_{000}) + \eta x^*} \right\}, \quad (5)$$

where $G(a, b) = F^{-1}(a) - F^{-1}(b)$.

Let $D_n = \{(x_i, y_i, z_i, \delta_i),\ i = 1, \ldots, n\}$ be the data after enrolling n patients in the trial. The likelihood function under the reparametrization is

$$\begin{aligned}
&L(\rho_{000}, \rho_{100}, \rho_{101}, \rho_{010}, \eta \mid D_n) \\
&\quad = \prod_{i=1}^{n} \left(H(\rho_{000}, \rho_{100}, \rho_{101}, \rho_{010}, \eta; x_i, y_i, z_i)\right)^{\delta_i} \\
&\qquad \times \left(1 - H(\rho_{000}, \rho_{100}, \rho_{101}, \rho_{010}, \eta; x_i, y_i, z_i)\right)^{1-\delta_i}
\end{aligned} \quad (6)$$

where

$$\begin{aligned}
H&(\rho_{000}, \rho_{100}, \rho_{101}, \rho_{010}, \eta; x_i, y_i, z_i) = F\Big(F^{-1}(\rho_{000}) \\
&+ \left(F^{-1}(\rho_{100}) - F^{-1}(\rho_{000})\right) x_i \\
&+ \left(F^{-1}(\rho_{010}) - F^{-1}(\rho_{000})\right) y_i \\
&+ \left(F^{-1}(\rho_{101}) - F^{-1}(\rho_{100})\right) z_i + \eta x_i y_i\Big).
\end{aligned} \quad (7)$$

2.2. Prior and Posterior Distributions. We consider the priors $\rho_{100} \sim Beta(a_1, b_1)$, $\rho_{010} \sim Beta(a_3, b_3)$, $\rho_{101} \sim Beta(a_2, b_2)$, and conditional on ρ_{100}, ρ_{010}, $\rho_{000}/min(\rho_{100}, \rho_{010}) \sim Beta(a_0, b_0)$, and $\eta \sim Gamma(a, b)$ with mean $E(\eta) = a/b$ and variance $\operatorname{var}(\eta) = a/b^2$. As described in [25], vague *Beta* priors are achieved by taking $a_j = b_j = 1$, $j = 0, 1, 2, 3$ while a vague Gamma prior is chosen with mean of 21 and variance of 540. The posterior distribution is given by,

$$\begin{aligned}
&\pi(\rho_{000}, \rho_{100}, \rho_{101}, \rho_{010}, \eta \mid D_n) \\
&\quad \propto \prod_{i=1}^{n} \left(H(\rho_{000}, \rho_{100}, \rho_{101}, \rho_{010}, \eta; x_i, y_i, z_i)\right)^{\delta_i} \\
&\qquad \times \left(1 - H(\rho_{000}, \rho_{100}, \rho_{101}, \rho_{010}, \eta; x_i, y_i, z_i)\right)^{1-\delta_i} \\
&\qquad \times \pi(\rho_{000} \mid \rho_{100}, \rho_{010})\, \pi(\rho_{100})\, \pi(\rho_{101})\, \pi(\rho_{010})\, \pi(\eta).
\end{aligned} \quad (8)$$

We used JAGS [28] to sample from the posterior distribution.

2.3. Trial Design. The algorithm for dose escalation/deescalation is similar to one discussed in [11, 25] with the additional binary covariate information. It uses the EWOC principle [26] where at each stage of the trial, we seek a dose of one agent using the current posterior distribution of the MTD of the agent given the current dose of the other agent and the next patient's baseline covariate value. For instance, if agent A is held constant at level x, the dose of agent B is y such that the posterior probability that y exceeds the MTD of agent B given the dose of agent $A = x$ and covariate value $Z = z$ is bounded by a feasibility bound α. Cohorts of two patients are enrolled simultaneously receiving different dose combinations. Specifically, the design proceeds as follows.

(1) Let $D_2 = \{(x_1, y_1, z_1, \delta_1), (x_2, y_2, z_2, \delta_2)\}$ be the data from the first cohort of two patients such that each patient receives the same dose combination $(x_i, y_i) = (X_{min,A}, Y_{min,B}) = (0, 0)$ for $i = 1, 2$.

(2) In the second cohort of two patients, patient 3 receives dose (x_1, y_3) and patient 4 receives dose (x_4, y_2). If $z_3 = z_1$ or $z_3 = z_2$, y_3 is the αth percentile of $\pi(\Gamma_{B|A=x_1,Z=z_3} \mid D_2)$. Otherwise, patient 3 receives the minimum dose combination $(X_{min,A}, Y_{min,B}) = (0, 0)$. If $z_4 = z_1$ or $z_4 = z_2$, x_4 is the αth percentile of $\pi(\Gamma_{A|B=y_2,Z=z_4} \mid D_2)$. Otherwise, patient 4 receives the minimum dose combination $(X_{min,A}, Y_{min,B}) = (0, 0)$. In general, the first time a patient is assigned to a given group defined by the binary covariate z always receives the minimum dose combination $(X_{min,A}, Y_{min,B})$ no matter how many patients have been treated in the other group, as described in [20]. Here, $\pi(\Gamma_{B|A=x_1,Z=z_3} \mid D_2)$ is the posterior distribution of the MTD of agent B given that the level of agent A is x_1 and the baseline covariate value of patient 3 is z_3, given the data D_2. $\pi(\Gamma_{A|B=y_2,Z=z_4} \mid D_2)$ is defined similarly. $\Gamma_{B|A=x}$ and $\Gamma_{A|B=y}$ can be expressed in terms of $\rho_{000}, \rho_{100}, \rho_{101}$, and ρ_{010}.

(3) In the i-th cohort of two patients,

   (a) If i is even, patient $(2i - 1)$ receives dose (x_{2i-3}, y_{2i-1}) and patient $2i$ receives dose (x_{2i}, y_{2i-2}), where $\mathrm{y}_{2i-1} = \Pi^{-1}_{\Gamma_{B|A=x_{2i-3},Z=z_{2i-1}}}(\alpha \mid D_{2i-2})$ and $x_{2i} = \Pi^{-1}_{\Gamma_{A|B=y_{2i-2},Z=z_{2i}}}(\alpha \mid D_{2i-2})$. Here, $\Pi^{-1}_{\Gamma_{A|B=y,Z=z}}(\alpha \mid D)$ is the inverse cumulative distribution function of the posterior distribution, $\pi(\Gamma_{A|B=y,Z=z} \mid D)$.

   (b) Similarly, if i is odd, patient $(2i - 1)$ receives dose (x_{2i-1}, y_{2i-3}) and patient $2i$ receives dose (x_{2i-2}, y_{2i}), where $x_{2i-1} = \Pi^{-1}_{\Gamma_{A|B=y_{2i-3},Z=z_{2i-1}}}(\alpha \mid D_{2i-2})$ and $y_{2i} = \Pi^{-1}_{\Gamma_{B|A=x_{2i-2},Z=z_{2i}}}(\alpha \mid D_{2i-2})$.

(4) Repeat step (3), until n patients are enrolled in the trial subject to the following stopping rule.

If the αth percentile of $\pi(\Gamma_{A|B=y,Z=z} \mid D)$ or $\pi(\Gamma_{B|A=x,Z=z} \mid D)$ is less than 0 or greater than 1, the recommended dose for the next patient is 0 or 1, respectively. In steps (2) and (3) above, a dose escalation is further restricted to be no more than a prespecified fraction of the dose range of the corresponding agent.

Stopping Rule. It is sufficient to evaluate a stopping rule for safety at the minimum dose combination because of the partial ordering assumption. The probability of DLT of all doses for both agents will be higher than θ if the probability at the minimum dose is higher than θ.We stop enrollment to the trial if $P(P(\mathrm{DLT} \mid (x, y) = (0, 0), z) \geq \theta + \delta_1 \mid \mathrm{data}) > \delta_2$, i.e., if the posterior probability that the probability of DLT at the minimum available dose combination in the trial exceeds the target probability of DLT is high for $z = 0, 1$. The design parameters δ_1 and δ_2 are chosen to achieve desirable model operating characteristics. At the completion of the trial, an estimate of the MTD curve for $z = 0, 1$ is obtained using (5) as

$$\widehat{C}_z = \left\{ (x^*, y^*) \in [0,1]^2 : y^* = \frac{G(\theta, \widehat{\rho}_{000}) - (G(\widehat{\rho}_{100}, \widehat{\rho}_{000}))\, x^* - (G(\widehat{\rho}_{101}, \widehat{\rho}_{100}))\, z}{G(\widehat{\rho}_{010}, \widehat{\rho}_{000}) + \widehat{\beta}_4 x^*} \right\}. \tag{9}$$

where $G(a, b) = F^{-1}(a) - F^{-1}(b)$, $\widehat{\rho}_{000}, \widehat{\rho}_{100}, \widehat{\rho}_{101}, \widehat{\rho}_{010}$, and $\widehat{\eta}$ are the posterior medians given the data D_n.

3. Simulation Studies

3.1. Simulation Set-Up and Scenarios. We present four scenarios for the true MTD curves as shown in Figure 1. The first scenario (a) is a case where the two true MTD curves for two groups are parallel and close to the minimum doses with ρ_{010} and ρ_{100} equal to each other and slightly higher than θ; the second scenario (b) is a case where the two true MTD curves for two groups are parallel but very close to each other; the third scenario (c) is a case where two true MTD curves for two groups are not parallel, and the last scenario (d) is a case where the two true MTD curves are parallel but lie far apart from each other and close to the maximum doses with ρ_{010} and ρ_{100} equal and largely lower than θ.

In addition, toxicity responses are generated assuming four link functions allowing us to evaluate misspecification: (i) logistic, $F(u) = (1 + e^{-u})^{-1}$, (ii) probit, $F(\mathrm{u}) = \Phi(u)$, where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution, (iii) normal, $F(u) = \Phi(u/\sigma)$ with $\sigma = 2$, and (iv) complementary log-log, $F(u) = 1 - e^{-e^{u}}$, where the parameter values of μ, β, γ, λ, and η were selected in such a way that they all have the same true MTD curve.

For each scenario, 1000 trials were simulated with the logistic link function as the working model, the target probability of DLT is fixed at $\theta = 0.33$, the trial sample size is $n = 40$ patients with 20 patients in each group, $\delta_1 = 0.05$ and $\delta_2 = 0.8$. Vague priors for the parameters $(\rho_{000}, \rho_{100}, \rho_{101}, \rho_{010}, \eta)$ were chosen. A variable feasibility bound α was started from 0.25 and increased by 0.05 each time when we compute the dose for the next patient until α was reached 0.5 [29]. A dose escalation is restricted to be no more than 20% of the dose range of the corresponding agent.

3.2. Design Operating Characteristics. In order to assess the performance of this method when designing a prospective trial, we evaluate its operating characteristics by comparing the following three designs.

(i) Design using a covariate (WC): patients are accrued to the trial sequentially and the dose combinations given to the next cohort of patients are calculated assuming model (1).

(ii) Design ignoring the covariate (IC): patients are accrued to the trial sequentially and the dose combinations given to the next cohort of patients are calculated assuming model (1) without the covariate, i.e., as in [25].

(iii) Design using parallel trials (PT): in each group, patients are accrued to the trial sequentially and model (1) without the covariate is implemented in each group.

We assume that we have balanced groups given a fixed sample size in which it is possible to carry out two separate studies.

3.2.1. Safety and Efficiency. We evaluate operating characteristics introduced by Tighiouart et al. (2014, 2017) [11, 25]. Safety is assessed through the average percent of DLTs across all trials and the percent of trials that have a DLT rate exceeding $\theta + 0.1$.

Efficiency is assessed using an overall MTD estimate, pointwise average bias, and percentage of selection. The overall MTD estimate is based on all trials:

$$\overline{C}_z = \left\{ \left(x^*, y^*\right) \in [0,1]^2 : y^* = \frac{F^{-1}(\theta) - F^{-1}\left(\overline{\rho}_{000}\right) - \left(F^{-1}\left(\overline{\rho}_{100}\right) - F^{-1}\left(\overline{\rho}_{000}\right)\right) x^* - \left(F^{-1}\left(\overline{\rho}_{101}\right) - F^{-1}\left(\overline{\rho}_{100}\right)\right) z}{F^{-1}\left(\overline{\rho}_{010}\right) - F^{-1}\left(\overline{\rho}_{000}\right) + \overline{\eta} x^*} \right\}. \quad (10)$$

where $z = 0, 1$, $F(\cdot)$ is the logistic function and $\overline{\rho}_{000}, \overline{\rho}_{100}, \overline{\rho}_{101}, \overline{\rho}_{010}$, and $\overline{\eta}$ are the average posterior medians of the parameters $\rho_{000}, \rho_{100}, \rho_{101}, \rho_{010}$, and η from all 1000 trials, respectively.

The pointwise average relative minimum distance from the true MTD curve $C_{z,true}$ to the estimated MTD curve for $z = 0, 1$ is defined as

$$d_{z(x,y)} = m^{-1} \sum_{i=1}^{m} d_{z(x,y)}^{(i)} \quad (11)$$

wherein

$$d_{z(x,y)}^{(i)} = \operatorname{sign}\left(y' - y\right) \times \min_{\{(x^*,y^*):(x^*,y^*) \in C_{z,i}\}} \sqrt{(x - x^*)^2 + \left(y - y^*\right)^2} \quad (12)$$

for every point $(x, y) \in C_{z,true}$, y' is such that $(x, y') \in C_{z,i}$ for all $(x, y) \in C_{z,true}$, and $C_{z,i}$ is the estimated MTD curve with binary covariate z for trial i.

The percentage of selection for $z = 0, 1$ uses the differences defined in (12):

$$P_{z(x,y)} = m^{-1} \sum_{i=1}^{m} I\left(\left|d_{z(x,y)}^{(i)}\right| \leq p \Delta_z (x, y)\right) \quad (13)$$

where $\Delta_z(x, y)$ is the Euclidean distance between the minimum dose combination $(0, 0)$ and the point (x, y) on the true MTD curve for $z = 0, 1$ and $0 < p < 1$.

3.3. Results

3.3.1. Trial Safety. Table 1 shows that the overall average percent of DLTs is always less than $\theta = 0.33$ varying between 16.84% and 30.42% for the WC design, and 21.55% and 30.10% for the IC design across four scenarios. In the group with $Z = 0$, the average percent of DLTs varies between 5.63% and 21.68% for WC design, 2.89% and 19.07% for IC design, and 6.43% and 21.95% for PT design. Safety becomes a concern when $Z = 1$ for IC design because high values of average percent of DLT are observed (varying between 32.37% and 46.15%). On the other hand, the average percent of DLT for all scenarios goes between 28.04% and 39.17% for WC design, and 14.54% and 29.55% for PT design. These rates are similar when using the true and misspecified models. In addition, the highest value of the percent of trials with an excessive rate of DLT as defined by a DLT rate exceeding $\theta + 0.1$ is 0.1% for WC design, 0.4% for IC design, and 0.0% for PT design in the group with $Z = 0$ while this value is higher for $Z = 1$ is 31.5% for WC design, 65.8% for IC design, and 1.8% for PT design. Thus, we conclude that the methodology is safe for WC and PT designs, but not for IC design. The other three misspecified models are shown in Table S1.

3.3.2. Trial Efficiency. Figure 1 shows the true and estimated MTD curves for each group of patients under the four scenarios (a)-(d) when using the three studied designs. The estimated MTD curves were obtained using (10) and DLT responses were simulated using the logistic link function. The estimated MTD curves are fairly close to the true MTD curves when accounting for a significant baseline covariate (scenarios a, c, d) using the WC and PT designs. When ignoring the covariate, the estimated MTD curve tends to be in between the true MTD curves. This shows that when

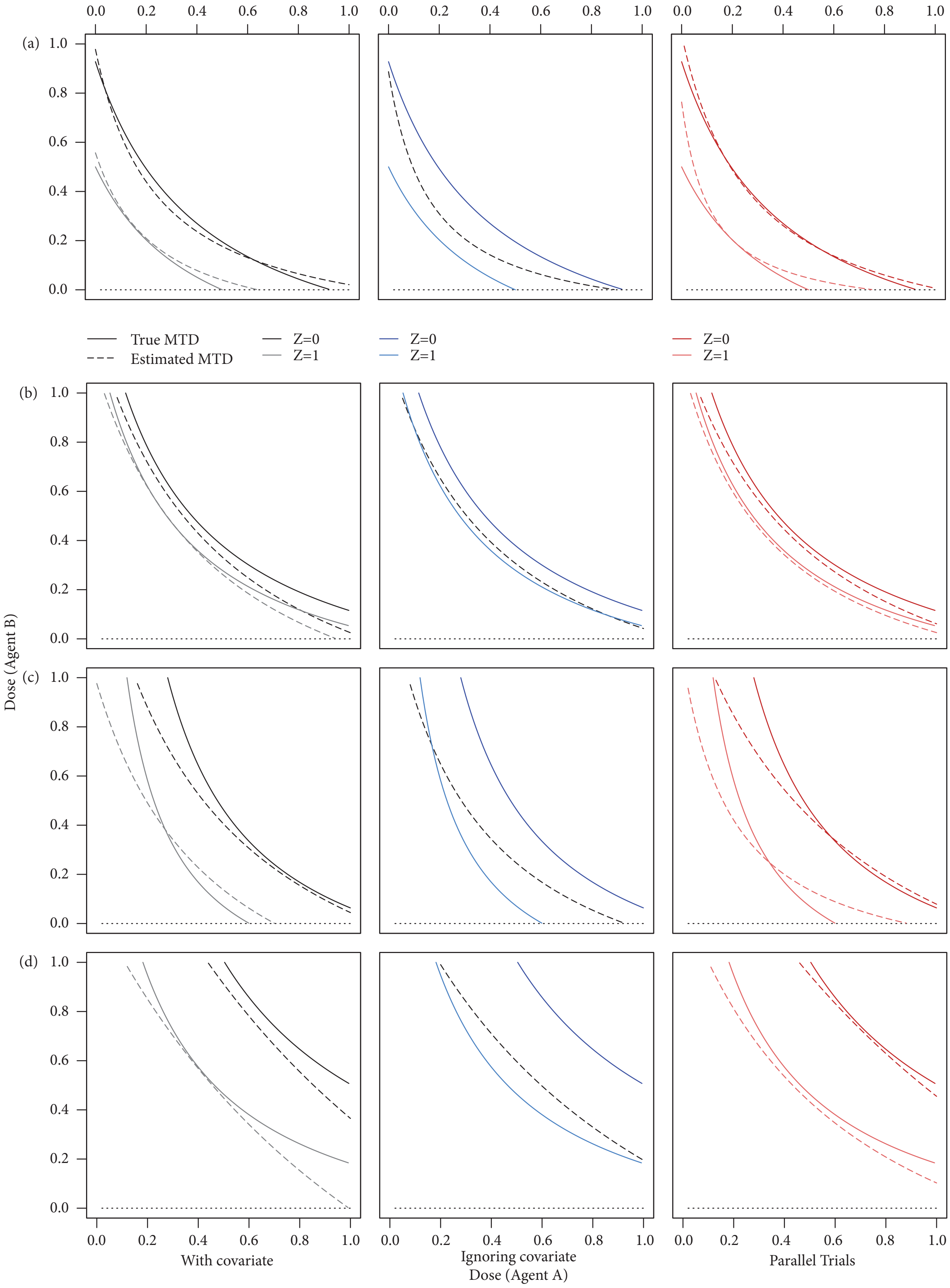


FIGURE 1: True and estimated MTD curves from $m = 1000$ simulated trials with designs using a covariate (WC), ignoring the covariate (IC), and parallel trials (PT) under four scenarios (a)-(d).

Table 1: Operating characteristics summarizing trial safety for designs using a covariate (WC), ignoring the covariate (IC), and parallel trials (PT) considering continuous dose combinations, $\theta = 0.33$.

Scenario	Design	Average % DLTs (% Trials: DLT rate < $\theta - 0.1$; % Trials: DLT rate > $\theta + 0.1$) Overall	 Z = 0	 Z = 1
(a)	WC	30.42 (7.1; 0.5)	21.68 (56.6; 0.0)	39.17 (0.4; 31.5)
	IC	30.10 (8.6; 0.4)	14.05 (86.7; 0.1)	46.15 (0.0; 65.8)
	PT	-	21.95 (56.6; 0.0)	29.55 (10.9; 1.8)
(b)	WC	24.34 (45.4; 0.0)	19.47 (73.8; 0.1)	29.21 (13.4; 1.8)
	IC	25.72 (34.6; 0.0)	19.07 (68.1; 0.4)	32.37 (10.2; 11.7)
	PT	-	16.87 (88.3; 0.0)	19.04 (78.0; 0.0)
(c)	WC	24.15 (47.7; 0.0)	13.98 (93.6; 0.0)	34.32 (2.7; 11.0)
	IC	24.90 (41.6; 0.0)	8.60 (97.3; 0.0)	41.21 (0.8; 42.6)
	PT	-	14.78 (96.5; 0.0)	22.21 (55.2; 0.1)
(d)	WC	16.84 (98.0; 0.0)	5.63 (100.0; 0.0)	28.04 (15.4; 0.0)
	IC	21.55 (73.3; 0.0)	2.89 (100.0; 0.0)	40.22 (1.0; 36.8)
	PT	-	6.43 (100.0; 0.0)	14.54 (97.6; 0.0)

Scenario ($\rho_{000}, \rho_{100}, \rho_{101}, \rho_{010}, \eta$): (a) (0.01, 0.4, 0.8, 0.4, 10), (b) (0.005, 0.1, 0.2, 0.1, 10), (c) (0.005, 0.2, 0.7, 0.01, 10), and (d) (10^{-4}, 10^{-3}, 0.05, 10^{-3}, 10).

the two MTD curves are well separated, not accounting for a baseline covariate results in suboptimal MTD curve estimation for the group of patients with high tolerance and a too toxic MTD curve recommendation for the other group.

Figure 2 displays the pointwise average relative minimum distance from the true MTD curve to the estimated MTD curve as defined by (11) under the four scenarios (a)-(d) when the DLT responses are simulated from the true and the other three misspecified models, respectively. This is a measure of pointwise bias for the MTD estimate. In the first scenario (a), the maximum absolute pointwise bias is 0.101 for $Z = 0$ and 0.099 for $Z = 1$. For WC design, the pointwise bias is negligible for low dose combinations and increases as we move away from the minimum dose combination with higher values when $Z = 0$ then $Z = 1$. For PT design, the pointwise bias is almost constant when $Z = 0$, such that it is lower than for WC design at the edges of the MTD curve and presents U-shape when $Z = 1$ with higher values than for WC design at the minimum dose combination and the edges of the MTD curve. In scenario (b), the maximum absolute pointwise bias is 0.069 for $Z = 0$ and 0.066 for $Z = 1$. WC and PT designs show U-shape pointwise bias with higher values for WC design than for PT design as we increase the dose combinations in any direction when $Z = 0, 1$. In scenario (c), the maximum absolute pointwise bias is 0.181 for $Z = 0$ and 0.155 for $Z = 1$. For WC design, the pointwise bias is negative for low dose combinations and approximates to zero as we increase the dose combination in any direction; For PT design, the bias is negative for low dose combinations and becomes positive until reaching the same initial magnitude when $Z = 1$ and a plateau lower than the initial magnitude when $Z = 0$ as we increase dose combinations in any direction. In scenario (d), the maximum absolute pointwise bias is 0.21 for $Z = 0$ and 0.139 for $Z = 1$. WC and PT designs are similar to each other, with WC showing higher pointwise bias for doses at the edge of MTD curve when $Z = 1$. IC design presents higher pointwise bias than for WC and PT designs in all scenarios. The other three misspecified models are shown in Figure S1.

Figure 3 shows the pointwise percent of trials for which the minimum distance from the true MTD curve to the estimated MTD curve is no more than $(100 \times p)$% of the true MTD for $p = 0.2$ as defined by (13). This can be interpreted as the percent of MTD recommendation for a given tolerance p. Under the first scenario (a), the percent of trials with correct MTD recommendation within 20% of the true value of the MTD varies between 62.6% and 99.9% using WC and PT designs, while it varies more widely between 28.6% and 100% for the IC design when the toxicities are generated from the true and misspecified models. The WC and PT design presents similar results to each other, with WC design showing slightly lower values than for PT design at the minimum dose combination. Under the second scenario (b), the percent of recommendation is similar between all designs varying between 84.9% and 97.6% for the WC design, 89.8% and 99.3% for the IC design, and 79% and 100% for the PT design. The WC design presents somewhat lower values than for IC design when $Z = 1$ at the minimum dose combination and at central part of the MTD curve when $Z = 0$. Under the scenario (c), the percent of recommendation is between 69.0% and 95.1% for the WC design, 67.1% and 95.3% for the PT design, while it is between 15.6% and 98.5% for IC design. The IC design is notably worse than WC and PT designs, except at the minimum dose combination when $Z = 1$; The percent of recommendation is always lower for PT design than for WC design when $Z = 0$ and at the edges of the MTD curve when $Z = 1$. In scenario (d), the percent of recommendation varies between 68.7% and 95.5% for WC design, 88.2% and 98.0% for the PT design, 50.7% and 89.2% for the IC design. As it was observed in the other scenarios, IC design performs worse than WC and PT designs. The PT design presents higher values than for WC design at the minimum dose combination when $Z = 0$ and at the edges of

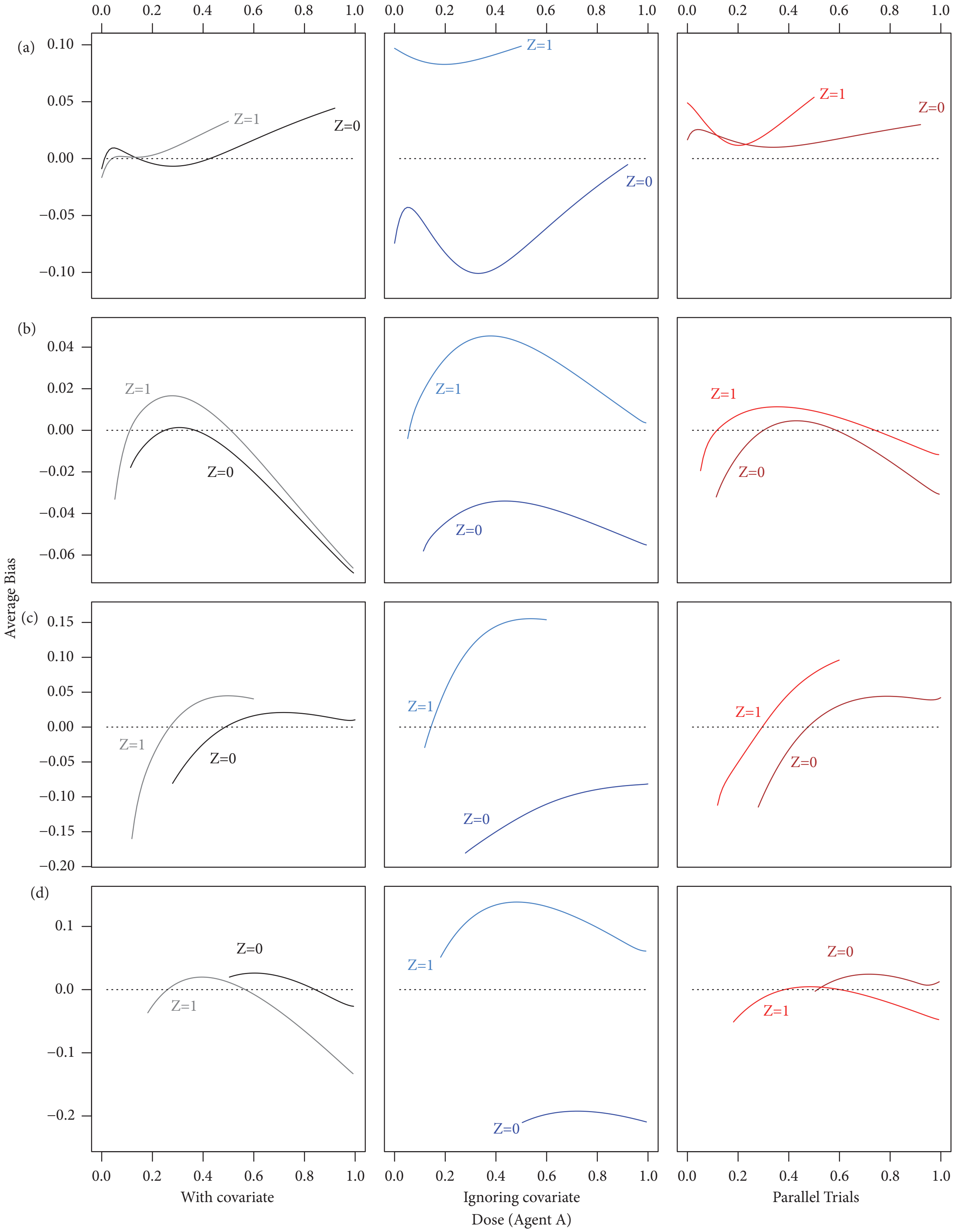


FIGURE 2: Pointwise average relative minimum distance from the true MTD curve to the estimated MTD curve with designs using a covariate (WC), ignoring the covariate (IC), and parallel trials (PT) under scenarios (a)-(d).

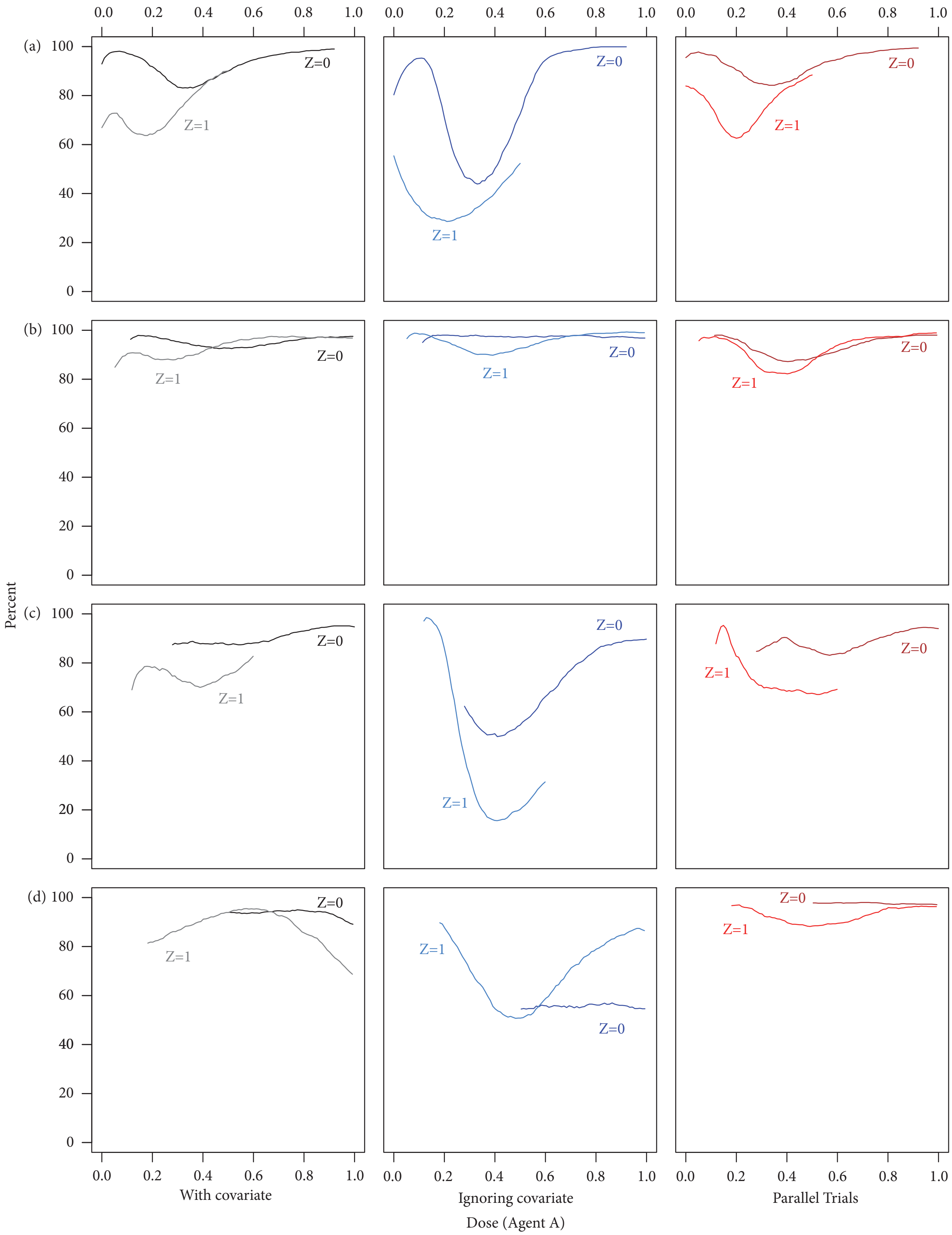


FIGURE 3: Pointwise percent of MTD recommendation for $p = 0.2$ with designs using a covariate (WC), ignoring the covariate (IC), and parallel trials (PT) under scenarios (a)-(d).

Table 2: A selected dose limiting toxicity scenario with $\theta = 0.33$ for $Z = 0, 1$ considering discrete dose combinations. True MTDs are shown in bold.

Dose level	Z = 0					Z = 1				
	1	2	3	4	5	1	2	3	4	5
5	0.25	**0.33**	0.40	0.48	0.70	0.45	0.53	0.60	0.68	0.90
4	0.20	0.26	**0.33**	0.43	0.55	0.40	0.46	0.53	0.63	0.75
3	0.13	0.16	0.24	**0.33**	0.39	**0.33**	0.36	0.44	0.53	0.59
2	0.05	0.13	0.18	0.28	**0.33**	0.25	**0.33**	0.38	0.48	0.53
1	0.001	0.05	0.13	0.20	0.27	0.201	0.25	**0.33**	0.40	0.47

the MTD curve when $Z = 0, 1$. The other three misspecified models are shown in Figure S2.

4. Discrete Dose Combinations

In this section, we show how the proposed methodology can be applied to a prespecified discrete set of dose combinations.

4.1. Approach. Let $(x_1, \ldots, x_r)$ and $(y_1, \ldots, y_s)$ be the doses of agents A and B, respectively. Following the notation of Section 2.1, $X_{min,A} = x_1$, $Y_{min,B} = y_1$, $X_{max,A} = x_r$, $Y_{max,B} = y_s$, the doses are standardized to be in the interval $[0, 1]$, and z is a binary baseline covariate. Trial design proceeds using the algorithm described in Section 2.3 where the continuous doses recommended in steps (2) and (3) are rounded to the nearest discrete dose levels. At the end of the trial, a discrete set Γ of dose combinations satisfying (i) and (ii) below is selected as MTDs. Let $C_{z,i}$ be the estimated MTD curve for $z = 0, 1$ at the end of the trial and denote by $d_z((x_j, y_k), C_{z,i})$ the Euclidean distance between the dose combination (x_j, y_k) and $C_{z,i}$ for $z = 0, 1$ as in (12).

(i) Let $\Gamma_{z,A} = \bigcup_{t=1}^{r}\{(x_t, y) : y = \operatorname{argmin}_{y_j} d((x_t, y_j), C_{z,i})\}$, $\Gamma_{z,B} = \bigcup_{t=1}^{s}\{(x, y_t) : x = \operatorname{argmin}_{x_j} d((x_j, y_t), C_{z,i})\}$, and $\Gamma_{z,0} = \Gamma_{z,A} \cap \Gamma_{z,B}$.

(ii) Let $\Gamma_z = \Gamma_{z,0} \backslash \{(x^*, y^*) : P(|P(DLT|(x^*, y^*), z) - \theta| > \delta_1 | D_n) > \delta_2\}$.

In (i), dose combinations closest to the MTD are selected by first minimizing the distances across the levels of drug A, then across the levels of drug B. In (ii), we exclude MTDs from (i) that either likely to be too toxic or too low. The design parameter δ_1 is selected after consultation with a clinician and the parameter δ_2 is selected after exploring a large number of scenarios for a given prospective trial. Following Tighiouart (2017) [25], $\delta_1 = 0.1$, $\delta_2 = 0.7$.

4.2. Operating Characteristics. The performance of the method is evaluated by calculating the percent of MTDs selection introduced in Tighiouart (2017) [25] estimating the probability that for a given scenario, a prospective trial will recommend a set of dose combinations that are all MTDs:

$$\mathrm{PS}_z = 100 \times \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left(\Gamma_{z,i} \subset \Gamma_{z,\delta}\right), \tag{14}$$

for $z = 0, 1$, where $\Gamma_{z,\delta} = \{(x_i, y_j) : |P(DLT|(x_i, y_j), z) - \theta| < \delta\}$ is the set of true MTDs such that the threshold parameter δ is fixed by a clinician. In the same way, the percent of selection at least K dose combinations that are MTDs is

$$\mathrm{PS}_z - \mathrm{K} = 100 \times \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left(\left|\Gamma_{z,i} \cap \Gamma_{z,\delta}\right| \geq K\right), \tag{15}$$

for $z = 0, 1$. In addition, the weighted average proportion of the recommended set of dose combinations which are MTDs discussed in [30] is given by

$$\mathrm{S}_{\Gamma_\delta} = \frac{\sum_{i=1}^{m} \left|\Gamma_{z,i} \cap \Gamma_{z,\delta}\right|}{\sum_{i=1}^{m} \left|\Gamma_{z,i}\right|}, \tag{16}$$

for $z = 0, 1$.

4.3. Illustration. We present one scenario as shown in Table 2 with $r = s = 5$ and the target probability of DLT is $\theta = 0.33$. We simulated $m = 1000$ trials using the sample size of $n = 40$ patients with 20 patients per group, and the same vague priors for ρ_{000}, ρ_{100}, ρ_{101}, ρ_{010} and η from Section 3 to compare the three designs with a covariate, ignoring the covariate, and parallel trials.

Table 3 shows that the overall average DLT is 25.1% for the WC design and 24.5% for the IC design. In the group with $Z = 0$, it is always far lower than θ and close to θ for the group with $Z = 1$. The percent of trials with an excessive DLT rate is not noticeable for all designs where the highest values are observed when using the IC design.

Table 4 shows that the design using parallel trials has highest values for the percent of MTDs selection (PS), percent of selection of at least 3 dose combinations (S-3), 2 dose combinations (S-2), and 1 dose combination (S-1), and weighted average percent of the recommended set of dose combinations ($\mathrm{S}_{\Gamma_\delta}$) statistics in the group with $Z = 0, 1$. The IC design shows the lowest values for all operating characteristics in both groups; PT design presents smaller values than for WC design when $Z = 0$, while shows higher values than for WC design when $Z = 1$, except for $\mathrm{PS}_z - 3$.

5. Conclusion

We described Bayesian adaptive designs for cancer phase I clinical trials using two drugs with continuous dose levels in the presence of a binary baseline covariate. The goal is to estimate the MTD curve in the two-dimensional Cartesian plane for a patient's specific baseline covariate value. The methodology extends the single agent trial design with a baseline covariate and two agents design without a covariate. In each case, vague priors were used to quantify the toxicity profile of each agent a priori. We used an algorithm for dose escalation where cohorts of two patients are enrolled simultaneously and the patients receive different dose combinations. We studied design operating characteristics of the method under four practical scenarios by comparing this method with the design that ignores the baseline covariate and design using parallel trials. In all simulations, we used a sample size of $n = 40$ patients, 20 patients in each group. We

Table 3: Operating characteristics summarizing trial safety for designs using a covariate (WC), ignoring the covariate (IC), and parallel trials (PT) considering discrete dose combinations.

Design	Average % DLTs (% Trials: DLT rate < $\theta - 0.1$; % Trials: DLT rate > $\theta + 0.1$)		
	Overall	Z = 0	Z = 1
WC	25.9 (31.9; 2.0)	17.3 (78.5; 3.0)	34.5 (7.2; 17.4)
IC	24.4 (48.0; 0.0)	14.2 (85.6; 0.0)	34.6 (7.7; 17.5)
PT	-	15.0 (91.1; 0.0)	28.1 (24.0; 5.0)

Table 4: Operating characteristics summarizing trial efficiency for $Z = 0, 1$ with designs using covariate (WC), ignoring covariate (IC), and parallel trials (PT) considering discrete dose combinations.

Covariate	Design	PS	$PS_z - 3$	$PS_z - 2$	$PS_z - 1$	S_{Γ_δ}
Z = 0	WC	47.3	35.3	51.3	80.1	66.3
	IC	19.6	14.9	23.0	51.6	37.9
	PT	46.2	9.9	33.6	58.3	56.4
Z = 1	WC	55.3	25.7	72.5	85.5	71.6
	IC	47.6	17.9	59.5	79.6	62.7
	PT	81.6	16.2	71.4	88.0	88.9

found that in general, the methodology is safe in terms of the probability that a prospective trial will result in an excessively high number of DLTs when accounting for a significant covariate. We used several measures to assess the efficiency of the estimate of the MTD. In the presence of a practically significant baseline covariate, the design with a covariate had a smaller pointwise average bias and a higher percent of MTD recommendation relative to a design which ignores the covariate and similar performance to parallel trials when the groups were balanced. When the two true MTD curves are very close, including a baseline covariate in the model results in a slightly higher but still negligible bias and a small reduction in percent of MTD recommendation relative to the design that ignores this covariate. Therefore, we stand to lose little if we include a practically not important covariate in the model. We further showed how this methodology is adapted to the discrete dose combinations and proposed statistics estimating the probability that a prospective trial will recommend a set of dose combinations that are all MTDs for a given scenario. The statistics are used in evaluating the performance of the proposed design with a covariate as compared to other designs ignoring the covariate and using parallel trials.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

An earlier version from the manuscript has been presented as poster in the Joint Statistical Meeting in Chicago, 2016 [31].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was partially funded by grants from the National Institute of Health, Grant Number 1R01CA188480-01A1 (Mourad Tighiouart, Sungjin Kim), and the National Center for Research Resources, Grant UL1RR033176, and is now at the National Center for Advancing Translational Sciences, Grant UL1TR000124 (Mourad Tighiouart), and P01 CA098912 (Mourad Tighiouart, Márcio Augusto Diniz).

Supplementary Materials

(i) Table S1: operating characteristics summarizing trial safety for designs using a covariate (WC), ignoring the covariate (IC), and parallel trials (PT) considering continuous dose combinations, $\theta = 0.33$. (ii) Figure S1: pointwise average relative minimum distance from the true MTD curve to the estimated MTD curve with designs using a covariate (WC), ignoring the covariate (IC), and parallel trials (PT) under scenarios (a)-(d). (iii) Figure S2: pointwise percent of MTD recommendation for $p = 0.2$ with designs using a covariate (WC), ignoring the covariate (IC), and parallel trials (PT) under scenarios (a)-(d). *(Supplementary Materials)*

References

[1] J. S. Lopez and U. Banerji, "Combine and conquer: Challenges for targeted therapy combinations in early phase trials," *Nature Reviews Clinical Oncology*, vol. 14, no. 1, pp. 57–66, 2017.

[2] P. F. Thall, R. E. Millikan, P. Mueller, and S.-J. Lee, "Dose-finding with two agents in Phase I oncology trials," *Biometrics: Journal of the International Biometric Society*, vol. 59, no. 3, pp. 487–496, 2003.

[3] K. Wang and A. Ivanova, "Two-dimensional dose finding in discrete dose space," *Biometrics: Journal of the International Biometric Society*, vol. 61, no. 1, pp. 217–222, 2005.

[4] G. Yin and Y. Yuan, "A latent contingency table approach to dose finding for combinations of two agents," *Biometrics: Journal of the International Biometric Society*, vol. 65, no. 3, pp. 866–875, 2009.

[5] G. Yin and Y. Yuan, "Bayesian dose finding in oncology for drug combinations by copula regression," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 58, no. 2, pp. 211–224, 2009.

[6] T. M. Braun and S. Wang, "A hierarchical Bayesian design for phase 1 trials of novel combinations of cancer therapeutic

agents," *Biometrics: Journal of the International Biometric Society*, vol. 66, no. 3, pp. 805–812, 2010.

[7] N. A. Wages, M. R. Conaway, and J. O'Quigley, "Continual reassessment method for partial ordering," *Biometrics: Journal of the International Biometric Society*, vol. 67, no. 4, pp. 1555–1563, 2011.

[8] N. A. Wages, M. R. Conaway, and J. O'Quigley, "Dose-finding design for multi-drug combinations," *Clinical Trials*, vol. 8, no. 4, pp. 380–389, 2011.

[9] M. J. Sweeting and A. P. Mander, "Escalation strategies for combination therapy Phase i trials," *Pharmaceutical Statistics*, vol. 11, no. 3, pp. 258–266, 2012.

[10] Y. Shi and G. Yin, "Escalation with overdose control for phase I drug-combination trials," *Statistics in Medicine*, vol. 32, no. 25, pp. 4400–4412, 2013.

[11] M. Tighiouart, S. Piantadosi, and A. Rogatko, "Dose finding with drug combinations in cancer phase I clinical trials using conditional escalation with overdose control," *Statistics in Medicine*, vol. 33, no. 22, pp. 3815–3829, 2014.

[12] M.-K. Riviere, Y. Yuan, F. Dubois, and S. Zohar, "A Bayesian dose-finding design for drug combination clinical trials based on the logistic model," *Pharmaceutical Statistics*, vol. 13, no. 4, pp. 247–257, 2014.

[13] A. P. Mander and M. J. Sweeting, "A product of independent beta probabilities dose escalation design for dual-agent phase I trials," *Statistics in Medicine*, vol. 34, no. 8, pp. 1261–1276, 2015.

[14] R. K. Ramanathan, M. J. Egorin, C. H. M. Takimoto et al., "Phase I and pharmacokinetic study of imatinib mesylate in patients with advanced malignancies and varying degrees of liver dysfunction: A study by the national cancer institute organ dysfunction working group," *Journal of Clinical Oncology*, vol. 26, no. 4, pp. 563–569, 2008.

[15] T. B. Leal, S. C. Remick, C. H. Takimoto et al., "Dose-escalating and pharmacological study of bortezomib in adult cancer patients with impaired renal function: A National Cancer Institute Organ Dysfunction Working Group Study," *Cancer Chemotherapy and Pharmacology*, vol. 68, no. 6, pp. 1439–1447, 2011.

[16] T. Satoh, T. Ura, Y. Yamada et al., "Genotype-directed, dose-finding study of irinotecan in cancer patients with UGT1A1*28 and/or UGT1A1*6 polymorphisms," *Cancer Science*, vol. 102, no. 10, pp. 1868–1873, 2011.

[17] P. M. LoRusso, K. Venkatakrishnan, R. K. Ramanathan et al., "Pharmacokinetics and safety of bortezomib in patients with advanced malignancies and varying degrees of liver dysfunction: Phase I NCI Organ Dysfunction Working Group Study NCI-6432," *Clinical Cancer Research*, vol. 18, no. 10, pp. 2954–2963, 2012.

[18] K.-P. Kim, H.-S. Kim, S. J. Sym et al., "A UGT1A1*28 and*6 genotype-directed phase i dose-escalation trial of irinotecan with fixed-dose capecitabine in Korean patients with metastatic colorectal cancer," *Cancer Chemotherapy and Pharmacology*, vol. 71, no. 6, pp. 1609–1617, 2013.

[19] J. O'Quigley, L. Z. Shen, and A. Gamst, "Two-sample continual reassessment method," *Journal of Biopharmaceutical Statistics*, vol. 9, no. 1, pp. 17–44, 1999.

[20] M. Tighiouart, G. Cook-Wiens, and A. Rogatko, "Escalation with overdose control using ordinal toxicity grades for cancer phase I clinical trials," *Journal of Probability and Statistics*, Art. ID 317634, 17 pages, 2012.

[21] J. O'Quigley and X. Paoletti, "Continual reassessment method for ordered groups," *Biometrics: Journal of the International Biometric Society*, vol. 59, no. 2, pp. 430–440, 2003.

[22] A. Ivanova and K. Wang, "Bivariate isotonic design for dose-finding with ordered groups," *Statistics in Medicine*, vol. 25, no. 12, pp. 2018–2026, 2006.

[23] Z. Yuan and R. Chappellb, "Isotonic designs for phase I cancer clinical trials with multiple risk groups," *Clinical Trials*, vol. 1, no. 6, pp. 499–508, 2004.

[24] M. R. Conaway, "A design for phase I trials in completely or partially ordered groups," *Statistics in Medicine*, vol. 36, no. 15, pp. 2323–2332, 2017.

[25] M. Tighiouart, Q. Li, and A. Rogatko, "A Bayesian adaptive design for estimating the maximum tolerated dose curve using drug combinations in cancer phase I clinical trials," *Statistics in Medicine*, vol. 36, no. 2, pp. 280–290, 2017.

[26] J. Babb, A. Rogatko, and S. Zacks, "Cancer phase I clinical trials: Efficient dose escalation with overdose control," *Statistics in Medicine*, vol. 17, no. 10, pp. 1103–1120, 1998.

[27] M. Tighiouart, Q. Li, S. Piantadosi, and A. Rogatko, "A Bayesian Adaptive Design for Combination of Three Drugs in Cancer Phase I Clinical Trials," *American Journal of Biostatistics*, vol. 6, no. 1, pp. 1–11, 2016.

[28] M. Plummer, "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd international workshop on distributed statistical computing*, vol. 124, p. 125, Vienna, 2003.

[29] M. Tighiouart and A. Rogatko, "Dose finding with escalation with overdose control (EWOC) in cancer clinical trials," *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, vol. 25, no. 2, pp. 217–226, 2010.

[30] M. A. Diniz, Q. Li, and M. Tighiouart, "Dose Finding for Drug Combination in Early Cancer Phase I Trials Using Conditional Continual Reassessment Method," *Journal of Biometrics Biostatistics*, vol. 8, 2017.

[31] S. Kim, M. A. Diniz, and M. Tighiouart, "A Bayesian Adaptive," in *Proceedings of the Design in Cancer Phase I Trials using Dose Combinations in the Presence of a Baseline Covariate*, A. S. Association, Ed., pp. 1336–1347, Alexandria, VA, 2016.

Hindawi
Journal of Probability and Statistics
Volume 2018, Article ID 1581979, 7 pages
https://doi.org/10.1155/2018/1581979

Review Article

Mixed Effects Models with Censored Covariates, with Applications in HIV/AIDS Studies

Lang Wu [1] and Hongbin Zhang [2]

[1] *Department of Statistics, University of British Columbia, Vancouver, Canada*
[2] *Department of Epidemiology and Biostatistics, Institute for Implementation Science in Population Health, City University of New York, USA*

Correspondence should be addressed to Lang Wu; lang@stat.ubc.ca

Received 26 March 2018; Accepted 5 May 2018; Published 3 June 2018

Academic Editor: Xu Zhang

Mixed effects models are widely used for modelling clustered data when there are large variations between clusters, since mixed effects models allow for cluster-specific inference. In some longitudinal studies such as HIV/AIDS studies, it is common that some time-varying covariates may be left or right censored due to detection limits, may be missing at times of interest, or may be measured with errors. To address these "incomplete data" problems, a common approach is to model the time-varying covariates based on observed covariate data and then use the fitted model to "predict" the censored or missing or mismeasured covariates. In this article, we provide a review of the common approaches for censored covariates in longitudinal and survival response models and advocate nonlinear mechanistic covariate models if such models are available.

1. Introduction

Mixed effects models are widely used in the analysis of clustered data, especially analysis of longitudinal data or survival data. In a longitudinal study, some variables are measured repeatedly over time, and these variables may be used either as responses or covariates, depending on study objectives. A common problem is that data on some of these variables may be left or right censored due to detection limits, may be missing at times of interest, or may be measured with errors. For example, in HIV/AIDS studies, viral load values may be left censored due to lower detection limits and may be missing or measured with substantial errors. In statistical analysis, these "incomplete data" issues must be addressed for correct statistical inference. In this article, we consider the case when these incompletely observed and time-varying variables are used as important covariates in mixed effects models for longitudinal response data or for time-to-event response data. To simplify the discussion, we focus on time-dependent covariates with left censoring, since similar methods/models may be used for right censoring or missing data or measurement errors in the covariates.

Longitudinal data with left censoring have received increasing attention in the literature in recent years (e.g., [1–8]). A common approach is to assume an empirical model for the covariate of interest based on the observed data, such as a linear mixed effects model. Then, the empirical model is used to "predict" the true covariate values when these values are censored, assuming the fitted model continues to hold for the unobserved censored values. A potential problem with this approach is that the assumed empirical covariate model based on the observed data may not hold for the censored covariate values, due to possibly different data-generation mechanisms for these "too small to observe" values. For example, in AIDS studies, censored viral loads below the detection limit may behave very differently from those above detection limit (observed values), due to a possibly different disease status for suppressed viral loads [6]. Moreover, the assumed model and distribution for the censored values cannot be verified based on observed data.

Recently, Kong and Nan [4] proposed an interesting approach based on ideas similar to that for right censored survival data, i.e., they used ideas similar to Cox models for right censored survival data for longitudinal data with left censoring. Yu et al. [6] proposed an approach which treats censored values as point mass. While these two approaches make no distributional assumptions for the censored values, the methods may not be efficient if censored values indeed follow a parametric distribution similar to that for the observed values.

In some applications such as HIV viral dynamics and pharmacokinetic modelling, mechanistic or scientific models can be derived based on the underlying *data-generation mechanisms*. These models are often *nonlinear* and are derived based on a set of differential equations which approximately describe the true data-generation mechanisms, so these models are justified biologically or scientifically (e.g., [9, 10]). Moreover, these mechanistic models have been shown to fit observed data quite well based on many data analyses [11]. Since these mechanistic models are based on underlying true data-generation mechanisms, they should hold for *censored values*, even though these values are not observed. Therefore, these models can be used to better "predict" the unobserved censored values than empirical models. In this article, we will provide a review of such approaches. The approaches are illustrated by an HIV/AIDS dataset.

2. Mixed Effects Models with Censored Covariates

In this section, we focus on generalized linear mixed effects models for longitudinal responses and survival models for time-to-event responses, with left censored and time-dependent covariates. The methods can be extended to other types of regression models in a conceptually straightforward way.

2.1. Generalized Linear Mixed Models with Censored Covariates. We first consider generalized linear mixed models (GLMMs) with a left censored and time-dependent covariate in a longitudinal study, following Zhang et al. [7]. Let y_{ij} be the response of interest measured for individual i at time t_{ij}, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, n_i$. Let x_{ij} be an important time-dependent covariate which is subject to left censoring, measurement errors, and missing data (assuming missing at random). We denote the unobserved true value of x_{ij} by x_{ij}^* in the presence of censoring or missing data or measurement errors. Let d be a known detection limit for x_{ij} such that x-values cannot be observed (detected) if $x_{ij} < d$ (i.e., left censoring), and let c_{ij} be the censoring indicator such that $c_{ij} = 1$ if $x_{ij} < d$ and $c_{ij} = 0$ otherwise. Let $\mathbf{z}_{ij}$ be a vector of other covariates.

Consider the following GLMM:

$$g\left(E\left(y_{ij}\right)\right) = x_{ij}^*\beta_1 + \mathbf{z}_{ij}^T\beta_2 + \mathbf{w}_{ij}^T\mathbf{a}_i, \qquad \mathbf{a}_i \sim N(0, A), \; i = 1, \ldots, n; \; j = 1, \ldots, n_i, \tag{1}$$

where $g(\cdot)$ is a known link function, β_j's are unknown parameters, $\mathbf{w}_{ij}$ is a subset of $(x_{ij}, \mathbf{z}_{ij})$, $\mathbf{a}_i$ contains random effects, and A is a unknown covariance matrix. We assume that the response y_{ij} follows a distribution in the exponential family such as a normal or Poisson or Binomial distribution. When the covariate x_{ij} is left censored or missing or measured with error, we may assume an empirical model for x_{ij} based on the observed x-data, such as a linear mixed effects (LME) model. Then we assume that the LME model continues to hold for censored or unobserved values and proceed for likelihood inference. However, as noted in Section 1, such an approach may be problematic since censored values may not follow the same model obtained based on the observed data.

When a mechanistic or scientific model is available for covariate x_{ij}, such as in HIV viral dynamics, the scientific model should hold not only for observed data but also for unobserved data (e.g., censored or mismeasured or missing data), so that the model can be used to provide better "predictions" for the unobserved true covariate values. Such a scientific model is often *nonlinear*. For longitudinal data with large between-individual variations, by introducing random effects in the nonlinear model to account for between-individual variations and within-individual correlations among repeated measurements, we obtain a nonlinear mixed effects (NLME) model. Thus, we assume that the covariate x_{ij} follows the following NLME model:

$$x_{ij} = h\left(t_{ij}, \mathbf{b}_i, \alpha\right) + e_{ij} \quad \left(\equiv x_{ij}^* + e_{ij}\right), \qquad \mathbf{b}_i \sim N(0, B), \; e_{ij} \text{ i.i.d.} \sim N\left(0, \sigma^2\right), \tag{2}$$

where $h(\cdot)$ is a known *nonlinear* function, vector $\mathbf{b}_i$ contains random effects, vector α contains fixed parameters, x_{ij}^* is the true covariate value at time t_{ij}, B is an unknown covariance matrix, and e_{ij}'s are random errors (measurement errors).

Note that when $h(\cdot)$ is a linear function (so model (2) is an LME model), the covariate model (2) is an empirical model which is chosen based on the observed covariate data. In a more general sense, the empirical models also include semiparametric or nonparametric mixed effects models. Such an empirical model is commonly used to address censoring, missing data, and measurement errors in the literature (e.g., [1, 2, 11]). When covariate x_{ij} is not normal, such as binary or count, generalized linear mixed models may be considered to fit observed covariate data, which are still empirical models. These empirical models may provide poor "predictions" to the unobserved data such as censored data.

2.2. Survival Models with Censored Time-Dependent Covariates. For survival models with time-dependent covariates, the covariates may also be left censored. Moreover, parameter estimation and inference for Cox models require that covariate values are available at event times [11]. However, this is usually not the case, since covariate values are unlikely to be available at all event times. Thus, this leads to missing covariate problems. The covariates may also be measured with errors, i.e., the observed covariate values may not be the true values but values with errors. In all cases, a common approach is to model the covariate process based on the observed covariate data and then use the fitted covariate model to "predict" the censored or missing covariate values. As noted in the previous section, a mechanistic or scientific covariate model may make better "predictions" than empirical covariate models, as shown in Zhang and Wu [8].

Here we consider a Cox model for the survival data with possible right censoring of the event times. For individual i, we define T_i to be the minimum of the observed event time T_i^* and the right censoring time C_i and define Δ_i to be the censoring indicator such that $\Delta_i = 1$ if the event time is right censored and $\Delta_i = 0$ otherwise, $i = 1, 2, \ldots, n$. Let $\lambda_i(t)$ be the hazard function for individual i at time t. The Cox model with time-dependent covariates can be written as

$$\lambda_i(t) = \lambda_0(t) \exp\left(\beta_1 x_i^*(t) + \boldsymbol{\beta}_2^T \mathbf{z}_i\right), \quad i = 1, 2, \ldots, n, \quad (3)$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \beta_2)^T$ is a vector of regression coefficients and $\lambda_0(t)$ is the (unspecified) baseline hazard function.

When the time-dependent covariate $x_i(t)$ is left censored or missing or measured with errors, inference for the Cox model can be challenging. Similar to the GLMM in the previous section, a common approach is to model the time-dependent covariate $x_i(t)$ based on observed covariate data, assuming the fitted covariate model holds for the censored covariate values. Again, such an empirical approach can be problematic if censored covariate values behave quite differently than observed values. The problem can be fixed if a mechanistic covariate model is available. We may again consider the mechanistic NLME model (2) to address censoring in the covariates.

3. Statistical Inference

For parameter estimation and inference, two methods are commonly used: the two-step method and the joint likelihood method. We briefly review the two methods below.

3.1. Two-Step Methods. To estimate the parameters in the models, a simple approach would be the so-called *two-step method*: in the first step we fit the covariate model based on the observed covariate data, and then in the second step we fit the response model *separately*, with the censored or missing covariate values substituted by their predicted values from the first step.

Specifically, consider the GLMM response model (1) and the covariate model (2). In the first step, we fit the NLME covariate model (2) to the observed covariate data and obtain estimates of the parameters $\widehat{\alpha}$ and the empirical Bayes estimates of the random effects $\widehat{\mathbf{b}}_i$. The predicted value of the covariate at time t is given by

$$\widehat{x}_i(t) = h\left(t, \widehat{\mathbf{b}}_i, \widehat{\alpha}\right). \quad (4)$$

Then, in the second step, we fit the following GLMM to the response data using the standard complete-data method for fitting GLMM

$$g\left(E\left(y_{ij}\right)\right) = \widehat{x}_{ij}\beta_1 + \mathbf{z}_{ij}^T\beta_2 + \mathbf{w}_{ij}^T\mathbf{a}_i. \quad (5)$$

If the covariate x value is censored or missing or mismeasured at time t_{ij}, its value is imputed by the predicted value $\widehat{x}_i(t_{ij}) = \widehat{x}_{ij}$.

An obvious issue with the above simple two-step method is that the estimation *uncertainty* in the first step is ignored in the second step. The standard error of the parameter estimate $\widehat{\beta}_1$ may be underestimated, leading to misleading inference for the parameter β_1. To fix this problem, we may use the bootstrap method to obtain more reliable standard errors of the parameters in the response model [11]. A parametric bootstrap method, which generates samples from the above fitted models, may be used to produce more reliable standard errors of the estimates. Still, the two-step method may not be efficient because covariate data and response data are not used simultaneously.

If the response data are survival data, the issues mentioned above for the two-step method remain. Moreover, in this case, the longitudinal covariate data may be truncated by the events such as death or dropouts. In this case, the two-step method may lead to biased estimation.

3.2. Joint Likelihood Method. A more desirable and formal method than the two-step method is to use the likelihood method based on the "joint likelihood" for both the response and covariates. Maximum likelihood estimates (MLEs) of all the unknown parameters in the two models may then be obtained *simultaneously* based on the joint likelihood for all observed data. If all assumed models and distributions hold, the MLEs are the most efficient estimates. Let $\boldsymbol{\theta}$ be the collection of all unknown parameters in the response and covariate models, and let $f(\cdot)$ denote a generic density function. The joint log-likelihood for the observed data is given by

$$l_{obs}(\theta) = \sum_{i=1}^{n} \log \int \prod_{j=1}^{n_i} f_y\left(y_{ij} \mid x_{ij}, \mathbf{a}_i; \beta\right) f_x\left(x_{ij} \mid \mathbf{b}_i, \alpha\right)^{1-c_{ij}} F_x\left(d \mid \mathbf{b}_i, \alpha\right)^{c_{ij}} f\left(\mathbf{a}_i, \mathbf{b}_i\right) d\mathbf{a}_i d\mathbf{b}_i \quad (6)$$

where $f_y(y_{ij} \mid x_{ij}, \mathbf{a}_i; \beta)$ is a density function from the exponential family, $F_x(d \mid \mathbf{b}_i, \alpha) = P(x_{ij} < d \mid \mathbf{b}_i, \alpha)$, and c_{ij} is the censoring indicator for the covariates.

Evaluation of the intractable integration in the log-likelihood $l_{obs}(\theta)$ can be computationally challenging, especially when the dimension of the random effects $(\mathbf{a}_i, \mathbf{b}_i)$ is higher. By treating the random effects $(\mathbf{a}_i, \mathbf{b}_i)$ as "missing data," we may use the EM algorithm to find the MLEs. Let $\mathbf{x}_{i,cen}$ be the censoring components of the covariate vector $\mathbf{x}_i$. By treating $(\mathbf{a}_i, \mathbf{b}_i, \mathbf{x}_{i,cen})$ as "missing data", Zhang et al. [7] proposed a Monte Carlo EM algorithm in which the E-step is implemented with a Gibbs sampler combined with rejection sampling methods. The Monte Carlo EM algorithm is still computationally intensive but is feasible. Alternatively, we may use computationally more efficient Laplace approximations or linearization methods to $l_{obs}(\theta)$ for approximate inference [11].

For the survival response models, the joint log-likelihood is given by

$$l^*_{obs}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} \int_{-\infty}^{d} \left[f\left(T_i, \Delta_i \mid \mathbf{a}_i; \boldsymbol{\alpha}, \boldsymbol{\lambda}_0, \boldsymbol{\beta}\right) \times f_x\left(x_{ij} \mid \mathbf{b}_i, \alpha\right)^{1-c_{ij}} F_x\left(d \mid \mathbf{b}_i, \alpha\right)^{c_{ij}} \times f\left(\mathbf{a}_i; \mathbf{A}\right) \right] d\mathbf{a}_i, \quad (7)$$

where

$$f\left(T_i, \Delta_i \mid \mathbf{a}_i; \boldsymbol{\alpha}, \boldsymbol{\lambda}_0, \boldsymbol{\beta}\right) = \left(\lambda_i\left(T_i \mid X_i^*\left(T_i\right); \boldsymbol{\alpha}, \boldsymbol{\lambda}_0, \boldsymbol{\beta}\right)\right)^{\Delta_i} \times S\left(T_i \mid X_i^*\left(T_i\right)\right), \quad (8)$$

with $S(t)$ the survival function defined as $S(t) = \exp(-\int_0^t \lambda(s)ds)$. Statistical inference can again be based on a Monte Carlo EM algorithm, although the computation can be more tedious due to the nonparametric baseline hazard in the Cox model.

4. Examples

In the following, we show two examples from an HIV/AIDS study. In the first example, we consider a Poisson generalized linear mixed model with censored covariates. In the second example, we consider a Cox survival model with censored covariates. In both examples, the time-dependent covariate is subject to left censoring and is modelled by a NLME model to address the censoring as well as missing data and measurement errors. The methods were implemented by Monte Carlo EM algorithms in R. R code is available upon request.

4.1. Generalized Linear Mixed Models with Censored Covariates. We consider an AIDS longitudinal dataset and study how viral load (VL) may relate to CD4 counts over time during an anti-HIV treatment. Viral loads usually have a lower detection limit so that viral load values below the limit cannot be observed, i.e., viral load may be left censored. Moreover, viral loads may be missing or measured with errors. As an illustration, we view CD4 count (y_{ij}) as the response and VL as a time-dependent covariate (x_{ij}), and we model the longitudinal CD4 counts as a Poisson GLMM:

$$\log\left(E\left(CD4_{ij}\right)\right) = \beta_{0i} + \beta_{1i}t_{ij} + \beta_{2i}VL^*_{ij} + \beta_3 TR_i + \beta_4 t_{ij} \times TR_i \quad (9)$$

where $\beta_{ki} = \beta_k + a_{ki}$, $k = 0, 1, 2$, a_{ki}'s are random effects, and TR denotes a treatment indicator. Since VL may be left censored and may be measured with errors, we consider the following mechanistic NLME model which is justified biologically [9, 10]:

$$VL_{ij} = \log_{10}\left(\alpha_{1i}e^{-\alpha_{2i}t_{ij}} + \alpha_{3i}e^{-\alpha_{4i}t_{ij}}\right) + e_{ij} \equiv VL^*_{ij} + e_{ij}, \quad (10)$$

where $\alpha_{ki} = \alpha_k + b_{ki}$, $k = 1, 2, 3, 4$, b_{ki}'s are random effects, and viral load values VL_{ij} are $\log_{10}$-transformed. The random effects are assumed to follow multivariate normal distributions with mean 0 and unstructured covariance matrices. As a comparison, we also fit observed VL data based on an empirical LME model (ELM):

$$VL_{ij} = \alpha_{1i} + \alpha_{2i}t_{ij} + \alpha_{3i}t^2_{ij} + \alpha_{4i}t^3_{ij} + e_{ij}. \quad (11)$$

The unknown parameters (α_k, β_k) are estimated using a Monte Carlo EM algorithm as described in Zhang et al. [7].

Figure 1 shows the NLME and ELM models fit to the observed viral loads of two randomly selected subjects, where the times are rescaled to be in $[0, 1]$. It suggests different fitted curves from the two covariate models. In particular, the predicted lines based on the NLME model fit the uncensored viral loads quite well; and for the censored portion, the lines follow the mechanistic model and preserve an overall nonlinear trend. On the other hand, the empirical LME model renders noticeable deviation of the fitted lines from the uncensored viral loads and imposes a linear or quadratic curve for the censored viral loads. Such discrepancies between the covariate model fitting, particularly in the censored portion, induce different parameter estimates in the response model. Table 1 summarizes the parameter estimates of the response CD4 model, with the covariate VL being fitted based on the NLME and ELM models, respectively. As we can see, the results of the parameter estimates are different. For example, the estimate of β_2, which measures the association between CD4 and VL, is significant at 5% level based on the NLME covariate model but is not significant based on the ELM covariate model. The results based on the NLME model should be more reliable since it provides more reliable predictions for the censored viral loads, since the NLME model may make better predictions for the unobserved censored values than the ELM method as the NLME model is based on the underlying data-generation mechanism which is the same for both observed and unobserved covariate values. The higher the percentage of censored/missing values is, the better the NLME model performs. This is confirmed by the simulation study in Zhang et al. [7].

4.2. Survival Models with Censored Covariates. As another example, we consider the foregoing dataset again, but now we focus on the occurrences of the first CD4:CD8 decline. The objective here is to determine if and how the time to the first CD4:CD8 decline may be related to treatment and viral load. We consider the following Cox survival model for the time to first CD4:CD8 decline:

$$\lambda_i(t) = \lambda_0(t) \exp\left(\beta_1 TR_i + \beta_2 VL^*_i(t)\right). \quad (12)$$

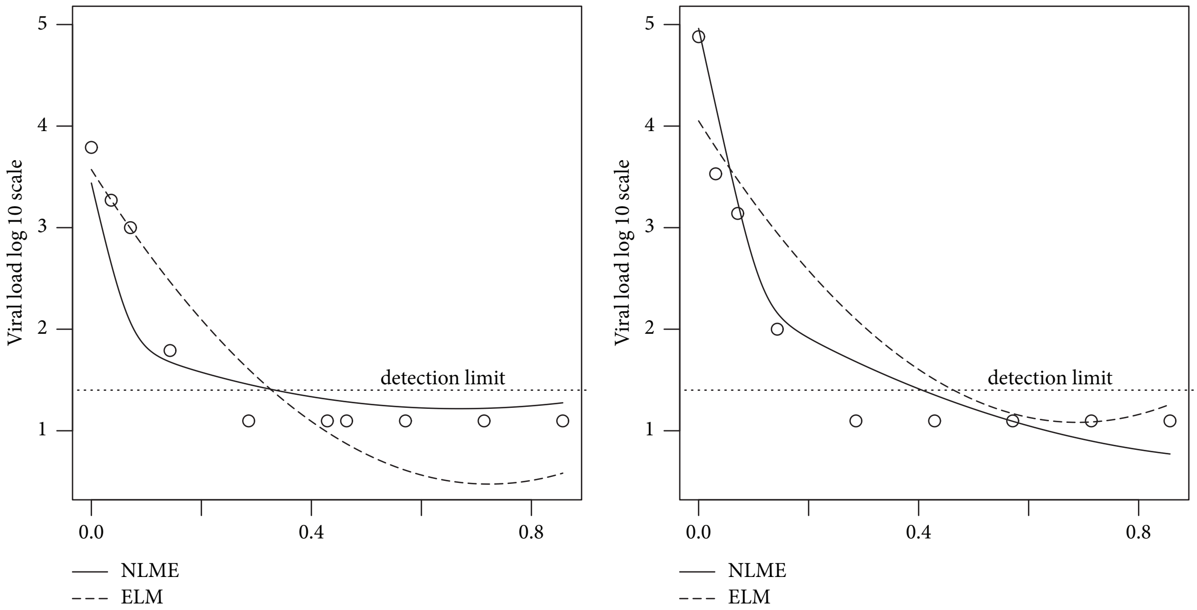


Figure 1: Fitted viral load curves for two randomly selected subjects. The open circles are the observed viral loads (the censored values are replaced by half the detection limit in $\log_{10}$ scale for simplicity). The solid line is the fitted curve based on the NLME model, while the dashed line is fitted curve based on the ELM model.

Table 1: Parameter estimates of the CD4 response model, based on the NLME and ELM covariate models respectively.

Response model parameter	NLME covariate model			ELM covariate model		
	Estimate	SE	p-value	Estimate	SE	p-value
β_0	5.98	0.16	0.00	6.15	0.23	0.00
β_1	−0.07	0.14	0.49	0.02	0.13	0.57
β_2	−0.19	0.06	0.01	−0.06	0.04	0.09
β_3	−0.71	0.59	0.19	−0.84	0.71	0.61
β_4	0.56	0.33	0.16	0.39	0.26	0.17

For this dataset, the Weibull distribution seems to provide a reasonable fit to the observed event times, so we consider the parametric Weibull distribution for the event times. For viral load, we use the same NLME and ELM models described in the first example.

Figure 2 shows, for two randomly selected subjects, the fitted lines to the observed viral loads based on the joint Cox survival model with the mechanistic NLME covariate model and empirical LME model (ELM), respectively, together with the corresponding estimated hazard functions and survival probability functions. We see that the mechanistic NLME model and the empirical LME model lead to different hazard and survival estimates. The NLME based joint model predicts monotonically increasing hazards, indicating the ever increasing risk of the event. On the other hand, the LME based model predicts more curved risk functions. Table 2 shows the results of the parameter estimates for the survival model. Here the differences seem relatively small, but as discussed, the predicted hazards and survival probabilities can be substantial. Since the NLME covariate model is derived based on reasonably biological justifications, they provide better "predictions" for censored (unobserved) viral loads and more reliable prediction for each individual's hazard and survival probability than the ELM covariate model, based on similar reasons as that for Table 1, which is also confirmed by simulations in Zhang and Wu [8].

5. Discussion

The nonlinear mechanistic covariate models are very appealing to address censoring and missing data in covariates, since the "predicted values" based on such models are more reliable than the commonly used empirical covariate models. These nonlinear mechanistic models are widely used in modelling HIV viral dynamics, pharmacokinetics, growth or decay, and some other areas [12, 13]. However, in many cases, such mechanistic models may not be available. In this case, an alternative approach is to treat the censored values as "point mass" to avoid unverifiable distributional assumptions for the censored values. The advantages of the nonlinear mechanistic covariate models are more obvious when the percentage of the censored values is higher, as confirmed in Zhang et al. [7]. The limitations of the nonlinear mechanistic covariate models are as follows: (i) in many

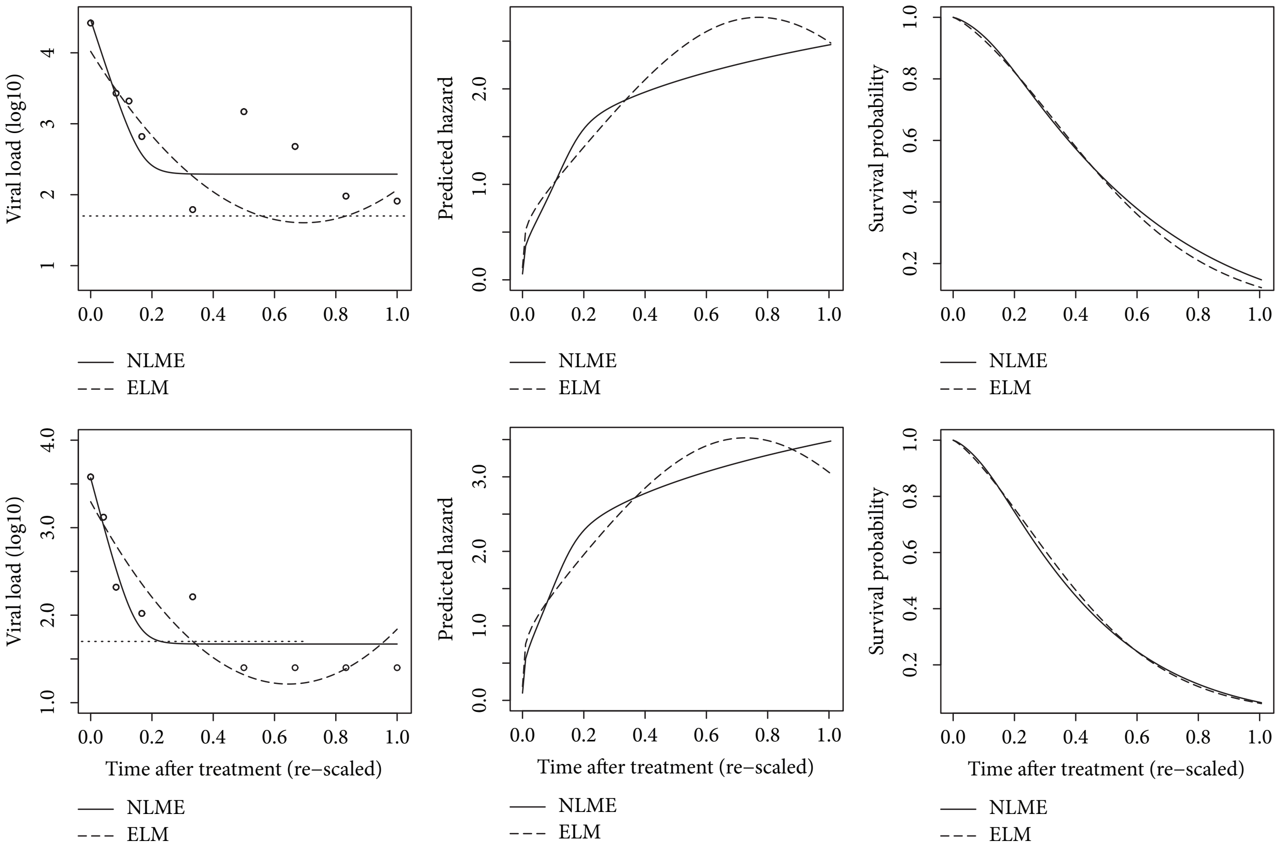


Figure 2: Plot of two individuals' (first row and 2nd row) fitted (predicted) viral load values and the corresponding hazard, survival functions based on the joint Cox and NLME model and empirical LME model (ELM), respectively. The open circles represent the observed viral loads. Left censored viral loads are replaced by one-half of the detection limit (in $\log_{10}$-scale).

Table 2: Parameter estimates in the Cox model based on the NLME and ELM covariate models respectively.

Cox model parameter	NLME covariate model			ELM covariate model		
	Estimate	SE	p-value	Estimate	SE	p-value
$\log(\lambda)$	1.60	1.05	0.10	1.44	1.13	0.60
$\log(\gamma)$	0.22	0.22	0.66	0.18	0.41	0.87
β_1	0.10	0.41	0.92	0.13	0.42	0.80
β_2	−0.40	0.54	0.39	−0.34	0.66	0.41

applications such mechanistic models may not be available and (ii) computation can be challenging, as discussed below.

Since the mechanistic covariate models are often nonlinear, computation is a main challenge in likelihood inference. Although Monte Carlo EM algorithms can almost always be used, they may offer potential problems such as very slow convergence or even nonconvergence. Moreover, the Monte Carlo EM algorithms usually need to be combined with Markov Chain Monte Carlo (MCMC) methods which are used to generate Monte Carlo samples in the E-step of the EM algorithms, making the computation even more challenging. When the dimensions of the random effects are high, we recommend approximate methods such as Laplace approximations and linearization methods as reviewed in Wu [11]. These approximate methods can be computationally much more efficient and provide reasonable approximations.

Data Availability

The dataset is available upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is partially supported by Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant no. 22R80742.

References

[1] P. W. Bernhardt, H. J. Wang, and D. Zhang, "Flexible modeling of survival data with covariates subject to detection limits via multiple imputation," *Computational Statistics & Data Analysis*, vol. 69, pp. 81–91, 2014.

[2] J. P. Hughes, "Mixed effects models with censored data with application to HIV RNA levels," *Biometrics*, vol. 55, no. 2, pp. 625–629, 1999.

[3] L. Wu, "A joint model for nonlinear mixed-effects models with censoring and covariants measured with error, with application to AIDS studies," *Journal of the American Statistical Association*, vol. 97, no. 460, pp. 955–964, 2002.

[4] S. Kong and B. Nan, "Semiparametric approach to regression with a covariate subject to a detection limit," *Biometrika*, vol. 103, no. 1, pp. 161–174, 2016.

[5] F. Vaida and L. Liu, "Fast implementation for normal mixed effects models with censored response," *Journal of Computational and Graphical Statistics*, vol. 18, no. 4, pp. 797–817, 2009.

[6] R. Fu and P. B. Gilbert, "Joint modeling of longitudinal and survival data with the Cox model and two-phase sampling," *Lifetime Data Analysis. An International Journal Devoted to Statistical Methods and Applications for Time-to-Event Data*, vol. 23, no. 1, pp. 136–159, 2017.

[7] H. Zhang, H. Wong, and L. Wu, "A mechanistic nonlinear model for censored and mismeasured covariates in longitudinal models, with application in AIDS studies," *Statistics in Medicine*, vol. 37, no. 1, pp. 167–178, 2018.

[8] H. Zhang and L. Wu, "A Mechanistic Nonlinear Model for Truncated and Mis-Measured Time-varying Covariates in Survival Models, with Applications in HIV/AIDS," *Journal of the Royal Statistical Society, C, accepted*, 2018.

[9] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho, "HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time," *Science*, vol. 271, no. 5255, pp. 1582–1586, 1996.

[10] H. Wu and A. A. Ding, "Population HIV-1 dynamics in vivo: Applicable models and inferential tools for virological data from AIDS clinical trials," *Biometrics*, vol. 55, no. 2, pp. 410–418, 1999.

[11] L. Wu, *Mixed effects models for complex data*, vol. 113 of *Monographs on Statistics and Applied Probability*, CRC Press, Boca Raton, FL, 2010.

[12] J. K. Lindsey, *Nonlinear models in medical statistics*, vol. 26 of *Oxford Statistical Science Series*, Oxford University Press, Oxford, 2001.

[13] M. Davidian and D. M. Giltinan, *Nonlinear Models for Repeated Measurements Data*, Chapman Hall, London, 1995.