Title
Flexible regression models for ROC and risk analysis, with or without a gold standard

Journal
Statistics in Medicine, 34(30)

ISSN
0277-6715

Authors
Branscum, Adam J
Johnson, Wesley O
Hanson, Timothy E
et al.

Publication Date
2015-12-30

DOI
10.1002/sim.6610

Peer reviewed

Research Article

Statistics in Medicine

Received 16 December 2014, Accepted 6 July 2015 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6610

Flexible regression models for ROC and risk analysis, with or without a gold standard

Adam J. Branscum,[a*†] Wesley O. Johnson,[b] Timothy E. Hanson[c] and Andre T. Baron[d]

A novel semiparametric regression model is developed for evaluating the covariate-specific accuracy of a continuous medical test or biomarker. Ideally, studies designed to estimate or compare medical test accuracy will use a separate, flawless gold-standard procedure to determine the true disease status of sampled individuals. We treat this as a special case of the more complicated and increasingly common scenario in which disease status is unknown because a gold-standard procedure does not exist or is too costly or invasive for widespread use. To compensate for missing data on disease status, covariate information is used to discriminate between diseased and healthy units. We thus model the probability of disease as a function of 'disease covariates'. In addition, we model test/biomarker outcome data to depend on 'test covariates', which provides researchers the opportunity to quantify the impact of covariates on the accuracy of a medical test. We further model the distributions of test outcomes using flexible semiparametric classes. An important new theoretical result demonstrating model identifiability under mild conditions is presented. The modeling framework can be used to obtain inferences about covariate-specific test accuracy and the probability of disease based on subject-specific disease and test covariate information. The value of the model is illustrated using multiple simulation studies and data on the age-adjusted ability of soluble epidermal growth factor receptor – a ubiquitous serum protein – to serve as a biomarker of lung cancer in men. SAS code for fitting the model is provided. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Bayesian semiparametric; disease diagnosis; medical test evaluation; Polya tree

1. Introduction

The development and statistical evaluation of screening, diagnostic, prognostic, and theragnostic procedures, such as imaging technologies and biomarker-based medical tests, are of great importance in public health and medical research. The ability of a biomarker – one based on continuous scale data – to distinguish diseased from healthy individuals is measured by the separation between distributions of test outcomes for the two groups. Because parametric models can fail to capture salient features of test outcome distributions, flexible statistical procedures for discriminant analysis of medical test data are at a premium [1–14]. We develop a general analytic framework that simultaneously handles two prominent tasks in the biosciences, namely to measure the performance of a continuous medical test and to determine an individual's likelihood of disease, all in the absence of training data based on a separate gold-standard (i.e., infallible) procedure. Although we focus on medical applications, the developed models and methods can be applied in other scientific fields.

Let D denote true disease status ($D = 1$ for disease present and $D = 0$ for disease absent, which for ease of discussion we refer to as 'healthy'). Let y denote a test outcome, or transformed outcome data from a medical procedure for diagnosing D. We refer to y as a 'test score', but in general, y can be any

[a]*Biostatistics Program, Oregon State University, Corvallis, Oregon 97331, U.S.A.*
[b]*Department of Statistics, University of California, Irvine, CA, U.S.A.*
[c]*Department of Statistics, University of South Carolina, Columbia, SC, U.S.A.*
[d]*Loring Life Sciences, Inc., Lexington, KY, U.S.A.*
[*]*Correspondence to: Adam J. Branscum, College of Public Health and Human Sciences, Oregon State University, Corvallis, Oregon 97331, U.S.A.*
[†]*E-mail: Adam.Branscum@oregonstate.edu*

continuous classifier. Without loss of generality, we adhere to the convention that larger values of y are associated with the presence of disease.

The statistical evaluation of test accuracy often involves estimating the receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC). The ROC curve illustrates the trade-off between a test's true positive and false positive probabilities across all possible cutoff thresholds $k \in \mathbb{R}$ that can be used to convert continuous test scores into dichotomous (positive or negative) outcomes. The true positive probability is the test's sensitivity of detection among diseased individuals, $Se(k) = \Pr(y > k \mid D = 1)$, and the false positive probability is $\Pr(y > k \mid D = 0) = 1 - Sp(k)$, where the specificity $Sp(k)$ is the proportion of healthy individuals who correctly test negative at cutoff k. The ROC curve is a plot of true positive versus false positive probabilities across k. The AUC is a summary measure of test accuracy that equals 0.50 for a test that correctly classifies disease status the same as a decision based on a fair coin toss and equals 1.0 for a perfect gold-standard test that always gives correct classification. When training data (i.e., gold-standard data) are available, and test outcome distributions are modeled parametrically, standard inferential methods apply (e.g., [15, Chapter 5]).

We focus on the more complicated task of estimation when test accuracy differs across covariates and when available data constitute a sample from a mixture of diseased and healthy individuals; the latter occurs when true disease status is unknown, which is an increasingly common obstacle that many studies designed to evaluate medical test accuracy must overcome (see [16] for a recent review). The methodology we develop has the flexibility to handle data from many different settings because (i) it allows for the inclusion of covariates related to test score distribution (we refer to these as 'test covariates') and covariates related to disease status (referred to as 'disease covariates'), (ii) the distributions of test scores for diseased and healthy populations are arbitrary and modeled using flexible mixture of finite Polya tree (MFPT) priors, and (iii) standard parametric ROC analysis and risk prediction, with or without gold-standard data, are available as special cases. In addition to developing a new flexible regression modeling framework and illustrating its use on simulated and real data, novel contributions of our research include theoretical results on identifiability and user-friendly SAS template code for fitting the model. To illustrate our new approach with and without gold-standard data, we consider a situation where training data were indeed collected, but where we analyze the data as if they were collected as a single sample with unknown disease status. We also analyze the data using the known disease status and compare with the no-gold-standard analysis. Similar illustrations in dichotomous medical test settings can be found in the works by Gastwirth and Johnson [17] and Johnson and Gastwirth [18].

1.1. Lung cancer data

We investigate the diagnostic potential of a soluble isoform of the epidermal growth factor receptor (sEGFR) present in blood as a biomarker for classifying lung cancer status in men. sEGFR is a member of the human epidermal growth factor receptor (*EGFR/HER/ERBB*) gene family, which has been linked to various human cancers, including lung, ovarian, and breast cancers (e.g., [19, 20]). The data were collected at the Mayo Clinic in Minnesota between 1998 and 2003 using a case-control design; there were 139 lung cancer cases and 88 controls. A goal of our analysis was to assess if and how test accuracy of sEGFR depends on age in men.

Several preliminary exploratory analyses were performed. Cases were split into two groups according to whether their age was below or above the median age (68 years) of cases in the sample. Two groups were similarly created for controls using their median age (37.5 years) as the cutoff. We found no clear evidence of a difference in the marginal distributions of $y = \ln(\text{sEGFR})$ for younger-aged versus older-aged cases (Wilcoxon $p = 0.18$). However, there was evidence of a difference (Wilcoxon $p = 0.0001$) in the distribution of y for controls based on age group, whereby older controls had higher serum sEGFR concentrations; the direction of the difference indicated that test accuracy might increase with increasing age. Moreover, with standardized age treated as a continuous predictor variable of y, a simple linear regression analysis using control data showed a significant association ($p < 0.0001$), while there was no evidence of a significant association using case data ($p = 0.13$). We thus included age as a (continuous) test covariate in our model for the control data.

We also ran a simple logistic regression analysis of disease status (lung cancer) on age and found that age was a predictor of case versus control status. Specifically, using the standardized age variable and independent $N(0, 1)$ priors for the two regression coefficients, the posterior median (95% posterior interval (PI)) for the odds ratio parameter that compares individuals who have standardized age of 0 with individuals whose age is one standard deviation above the mean was 8.9 (5.5, 15.3). Ideally, the study

would have matched on age, but because it did not, any analysis of these data should adjust for age. We thus regard age as a disease covariate. Finally, the Anderson–Darling test of normality gave $p < 0.0001$ for lung cancer cases (a kernel density estimate indicated a bimodal distribution for cases), $p = 0.06$ for younger controls, and $p = 0.19$ for older controls. As a result of all of these findings, we believe that these data are almost ideal for illustrating the need for our methods.

Caveat: Mixture modeling is difficult in general. In the two-group normal–normal setting, a mixture model with unknown mixing proportion is not identifiable unless it is assumed that one mean is smaller than the other, for example. In the more general case, with the two distributions being modeled nonparametrically, even the assumption of stochastic domination of one distribution over the other does not make the model identifiable. Moreover, in the area of medical classification with multiple binary tests, it is often the case that models either lack identifiability or require potentially strong assumptions in order to guarantee identifiability [21, 22]. The approach taken here buys identifiability based on having additional information, including continuous test outcomes instead of dichotomous outcomes and covariate information that should be helpful in mitigating the lack of a gold standard. The caveat here, however, is that the realized identifiability comes with the price of the assumption that the model for the relationship between disease status and disease covariates is at least approximately 'true'. We hedge our wording here because we agree with Box [23] that all models are wrong but some are useful. The essence of the caveat is that there is no free lunch. We either abandon statistical modeling of no-gold-standard data or we proceed with models like ours that ultimately rest on the ability to discern, in the absence of the benefit of the previous preliminary analysis, a set of covariates that are related to disease status and would thus be helpful in achieving identifiability. On the other side of the spectrum, our new method is a valuable contribution to Bayesian semiparametric ROC regression in the gold-standard case.

The remainder of the paper is organized as follows. Section 2 presents background on Polya trees and discusses previous methods for ROC curve estimation and risk prediction without a gold standard. Section 3 develops our general semiparametric regression model in the absence of gold-standard data and discusses several special cases. Methods are illustrated in Section 4 using simulated data and the sEGFR lung cancer data. Concluding remarks are given in Section 5, and theoretical results on model identifiability and SAS code are provided in the appendices.

2. Background

Standard frequentist and Bayesian approaches to test evaluation are catalogued in the texts by Pepe [24] and Broemeling [15], respectively; see also [25]. These texts focus primarily on parametric analysis of gold-standard data.

There is a large body of research on modeling binary medical test data without a gold standard. However, the literature on modeling continuous test scores without a gold standard is comparatively small. Frequentist parametric approaches have been developed by Henkelman, Kay, and Bronskill [26], Beiden *et al*. [27], and Kupinski *et al*. [28]. A Bayesian parametric approach was developed by Choi *et al*. [29], who used a bivariate two-group normal model for correlated tests in ROC analysis, while Choi, Johnson, and Thurmond [30] developed methods for risk prediction based on parametric models without a gold-standard test. Collins and Huynh [16] reviewed many frequentist and Bayesian methods for evaluating the accuracy of binary, ordinal, and continuous tests in the absence of a gold standard.

Hall and Zhou [31] developed a nonparametric approach for estimating the densities associated with data distributions for diseased and healthy populations, and nonparametric methods for ordinal tests were developed by Zhou, Castelluccio, and Zhou [32] with follow-up work that incorporated conditional dependence by Albert [33], all without gold-standard data. However, a limitation of these nonparametric procedures is the requirement of three or more tests to ensure identifiability. Wang *et al*. [2] developed a Bayesian multinomial model for grouped data in ROC studies involving two continuous tests; a related multinomial modeling approach was used by Fosgate, Scott, and Jordan [3]. A parametric Bayesian model for no-gold-standard data was developed by Wang *et al*. [34]; their group also developed methods for parametric analysis of longitudinal test score data [35], as did Norris, Johnson, and Gardner [36].

Bayesian nonparametric procedures for ROC analysis of gold-standard data have been developed using Dirichlet process mixtures [1, 5], dependent Dirichlet processes [13], finite or MFPTs [5, 37], and Gaussian processes [38]. Flexible methods based on multivariate mixtures of Polya trees to model gold-standard data from multiple continuous tests [6] and the bootstrap [7] have also been successful.

Pepe [39] and Rodríguez-Álvarez, Tahoces, and Cadarso-Suárez [40] considered the so-called 'induced semiparametric location-scale' models with test covariates, but no disease covariates. These models are not identifiable in the absence of a gold standard [31].

In contrast, Branscum *et al.* [4] developed a nonparametric model for continuous no-gold-standard data that uses flexible MFPT priors for the disease and healthy test score distributions and that relates disease covariates to latent disease status through a binomial regression model. With dichotomous tests, Magder and Hughes [41] and McInturff *et al.* [42] also used disease covariates, in part so that their models would be identifiable, but also because it was sensible to model disease prevalence as a smoothly varying function of certain covariates. The extension by Branscum *et al.* [4] of the model used by Magder and Hughes is also identifiable (under mild conditions), which we establish in Appendix A.

The new model we develop extends that of [4] to account for the possibility of test covariate information. Our model provides inference for covariate-specific ROC curves, AUC, and partial AUC, in addition to risk prediction of disease based on an individual's disease covariate vector ($\mathbf{x}^*$), test covariate vector ($\mathbf{x}$), and scalar test score (y). We expect the discriminatory ability of a test/biomarker to be greatly enhanced in many situations through the joint use of test and disease covariate information.

3. Methods

The data are $\{(\mathbf{x}_i, \mathbf{x}_i^*, y_i) : i = 1, \ldots, n\} \equiv (\mathbf{X}, \mathbf{X}^*, \mathbf{Y})$. The test covariates, $\mathbf{x}_i$, may be distinct from the disease covariates, $\mathbf{x}_i^*$, or they may overlap. For example, both might include age as in the lung cancer analysis. Continuous test scores (y_i) are obtained on n individuals from a population with overall disease prevalence π. Test scores are often transformed in parametric ROC analysis in order to conform to standard probability models. We denote the data as y_i regardless of whether or not they have been transformed.

In this section, we first discuss the use of Bayesian nonparametric Polya tree priors, then elucidate our semiparametric model for data analysis with and without a gold standard, and finally discuss how inferences are made. The modeling is the most interesting part and involves first modeling the latent disease status for individuals as independent Bernoulli random variables, and then conditional on disease status, modeling test outcomes. Test outcomes in the diseased and healthy populations are modeled by location-scale families of distributions that are assigned finite Polya tree priors. The model presented in the following can be easily adapted to handle a variety of scenarios, including the one needed for the lung cancer data.

3.1. Basics of MFPT priors

Consider the general case of modeling a continuous test outcome that varies according to an unknown distribution P. Because we want the outcome distribution to be essentially unconstrained, we do not force P to be a member of a parametric family, but rather, a family that encompasses a broad class of distributions. We allow the shape of P to be more-or-less arbitrary, thus allowing for unanticipated skewness and/or multimodality in the data.

Polya tree priors were discussed by Ferguson [43], with development for statistical modeling and extensions to mixtures of Polya trees by Lavine [44,45], Hanson and Johnson [46], and Hanson [47]. For a simple introduction to Polya trees, see Christensen *et al.* [48, Chapter 15] or Christensen, Hanson, and Jara [49]. Here, we give the basics of the MFPT prior distribution.

The first step in the definition of a Polya tree involves dyadic partitioning of the sample space. This involves a tree of partitions that begins with a partition into two sets, then each of those sets gets split to produce a finer partition of four sets, and so on. Corresponding to each pair of adjacent sets produced from a split is a pair of branch probabilities that must add to one. The collection of branch probabilities determines a particular P. Placing independent Dirichlet distributions on the pairs of branch probabilities results in a random probability distribution P called a Polya tree. The parameters of the Dirichlet distributions can be selected so that P is absolutely continuous with probability one, and also so that $E(P) = P_0$, a prior 'guess' for P. The parameters of the Dirichlet distributions are determined up to a positive constant, c, where large c leads to a prior for P that is concentrated around P_0 and small c leads to higher prior variability and provides a better opportunity for data-driven flexibility in making inferences. The dyadic partition is determined by P_0, so the only parameters of the Polya tree prior are P_0 and c. A finite Polya tree is obtained by truncating at some level, say J. We say that $P \sim \text{FPT}_J(P_0, c)$.

To enhance flexibility and smoothness, we replace the single P_0 with a family of parametric distributions, $\{P_\theta : \theta \in C\}$, for some appropriate set C. For example, P_θ might correspond to a normal family with unknown mean and variance. In this way, we center the prior on a parametric family, with the weight parameter c representing our prior confidence in that family as well as influencing the ability of the data to generate a posterior departure from the family. We complete this part of the model by placing a prior distribution on $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$, the same as we would if P were specified directly by the parametric family. The resulting marginal distribution for P is called a MFPT prior. It is common to describe them as 'nonparametric' because their flexibility produces robust inference; however, MFPT-based models are usually parametric with a high-dimensional parameter vector [49].

3.2. Semiparametric model

Let z_i denote latent disease status with $z_i = 1$ if subject i is diseased and $z_i = 0$ otherwise, and let π_i denote the probability that subject i is diseased. The test score data are modeled independently according to a mixture distribution with density $f_i(y) = (1-\pi_i)g_0(y) + \pi_i g_1(y - \mathbf{x}_i'\boldsymbol{\beta})$. In general, the model for healthy individuals can also depend on covariates, or, as in the case of the lung cancer data, the distribution of test scores for healthy individuals may depend on test covariates, while the model for data from the diseased population may not. With gold-standard data, the z_i's are known, and the data are modeled directly according to densities g_0 (healthy population) and g_1 (diseased population).

Let G_0 and G_1 denote the cumulative distribution functions (CDFs) associated with the densities g_0 and g_1, respectively. Under our sampling model, test scores for healthy individuals vary according to g_0, and a regression model characterizes the distribution of scores for diseased individuals. We present this particular model for concreteness, and because extensions and variations of it are straightforward. The model is thus specified as

$$\begin{aligned} z_i &\sim \text{Bernoulli}(\pi_i), \qquad \pi_i = F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}) \\ f(y_i \mid z_i) &= (1-z_i)g_0(y_i) + z_i g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \\ G_j &\sim \text{FPT}(G_{j\theta_j}, c_j),\ j = 0, 1, \qquad p(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}_0)p(\boldsymbol{\theta}_1)p(\boldsymbol{\alpha})p(\boldsymbol{\beta}), \end{aligned}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$, $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_s)'$, and F is a known CDF (for example, the standard normal or logistic). The distribution G_1 is (easily) constrained to have median zero in order to alleviate confounding between β_0 and the location of G_1 [46].

The parametric version of the model (when $c_0, c_1 \to \infty$), with or without disease covariates, has been termed a 'mixture of experts' model and is identifiable up to some order restrictions [50]. Our argument for identifiability of the semiparametric model is given in Appendix A. Notably, identifiability is achieved based primarily on the minor condition that one disease covariate regression coefficient be forced to be positive (or negative).

Although other distributions can be used, in our applications the parametric models $G_{0\theta_0}$ and $G_{1\theta_1}$ are normal distributions, namely $N(\mu_0, \sigma_0^2)$ and $N(0, \sigma_1^2)$, respectively, so that $\boldsymbol{\theta}_0 = (\mu_0, \sigma_0)'$ and $\boldsymbol{\theta}_1 = \sigma_1$. With gold-standard data, our model is thus centered at the covariate-adjusted binormal model considered by Faraggi [51]. The independent priors used for the means and standard deviations are $\mu_0 \sim N(a_{\mu_0}, b_{\mu_0}^2)$, $\sigma_0 \sim \text{Uniform}(a_{\sigma_0}, b_{\sigma_0})$, and $\sigma_1 \sim \text{Uniform}(a_{\sigma_1}, b_{\sigma_1})$, where the hyperparameters are fixed constants. It has become commonplace to set c_j equal to 1 or to model it with a prior distribution.

There are many approaches for using expert opinion or previous data to construct priors on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, including conditional means priors [52] and partial information priors for regression coefficients in generalized linear models [48, Chapter 8]. Depending on the setting and data available, opinions of experts can be pooled for use in analyzing latent disease status [53]. We have found that g-priors for $\boldsymbol{\alpha}$ in logistic regression [54] work well in a variety of simulations. In this approach, a multivariate normal g-prior for $\boldsymbol{\alpha}$ is constructed that corresponds to prior information about the probability of disease when averaged over the set of disease covariates that are included in the logistic regression model. Specifically, we first place an informative beta prior distribution on $\tilde{\pi}$, the (marginal) probability of disease when averaged over the distribution of disease covariates $H(d\mathbf{x}^*)$, that is, $\tilde{\pi} = \int \text{logit}^{-1}(\mathbf{x}^{*\prime}\boldsymbol{\alpha})H(d\mathbf{x}^*)$. Then, a g-prior on $\boldsymbol{\alpha}$ is obtained that induces a prior on $\tilde{\pi}$ that matches the elicited beta distribution by selecting appropriate values for g and b in $\boldsymbol{\alpha} \sim N_{s+1}(b\,\boldsymbol{e}, gn(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1})$, where $\mathbf{X}^* = [\mathbf{x}_1^* \cdots \mathbf{x}_n^*]'$ is the design matrix containing disease covariates. The last s elements of vector $\boldsymbol{e}$ are all zero, and its first element is one so that a prior mean of b is placed on α_0. Detailed methods and `R` code for obtaining g and b are described elsewhere [54], and example `SAS` code for calculating g and b is in Appendix B.

When prior information is lacking, prior distributions can be selected to be relatively diffuse. One approach in the case of a logistic regression model for relating disease status to disease covariates is to standardize all continuous covariates and place independent $N(0, 1)$ priors on their regression coefficients; induced priors on probabilities of disease tend to be rather diffuse with this selection [48, 55]. Simple methods for placing a multivariate normal g-prior on the vector of logistic regression coefficients in the absence of available prior information are also available [54]. It is common to use g-priors, diffuse normal, or even improper constant priors for the β_k's.

3.3. Model implementation and inference

SAS code that can be used as a template for fitting the mixture regression model is in Appendix B. The code can be easily modified to fit a variety of models using Gibbs sampling by iteratively simulating from the full conditional distributions of the parameters μ_0, σ_0, σ_1, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\Pi}_0$, and $\boldsymbol{\Pi}_1$, where $\boldsymbol{\Pi}_0$ and $\boldsymbol{\Pi}_1$ denote the collection of FPT branch probabilities in the priors on G_0 and G_1, respectively.

After burn-in, the simulated values from the Gibbs sampler are generated from the joint posterior distribution $p(\mu_0, \sigma_0, \sigma_1, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Pi}_0, \boldsymbol{\Pi}_1 \mid \mathbf{Y})$. The simulated iterates can be used to obtain posterior inferences for any parameter of interest, including covariate-specific ROC curves, $\text{ROC}(t \mid \mathbf{x})$, and predictive probabilities of disease for individuals with joint covariate vector $(\mathbf{x}^*, \mathbf{x})$ and scalar test score y, namely $\Pr(z = 1 \mid y, \mathbf{x}^*, \mathbf{x}, \mathbf{Y})$.

The covariate-specific ROC curve is given by

$$\text{ROC}(t \mid \mathbf{x}) = S_1\left(S_0^{-1}(t) \mid \mathbf{x}\right),$$

where $S_0(t) = 1 - G_0(t)$, $S_1(\cdot \mid \mathbf{x}) = 1 - G_1(\cdot \mid \mathbf{x})$, and $G_1(\cdot \mid \mathbf{x})$ denotes the CDF of $\mathbf{x}'\boldsymbol{\beta} + \epsilon$ when $\epsilon \sim G_1$. Hence, $G_1(v \mid \mathbf{x}) = G_1(v - \mathbf{x}'\boldsymbol{\beta})$, and therefore, $S_1(v \mid \mathbf{x}) = 1 - G_1(v - \mathbf{x}'\boldsymbol{\beta})$. Realizations at iteration j of the Gibbs sampler are used to obtain values for S_0 and S_1, and the corresponding iterate $\text{ROC}^{(j)}(t \mid \mathbf{x}) = S_1^{(j)}(S_0^{-1\,(j)}(t) \mid \mathbf{x})$. With m iterates, we obtain a numerical approximation to the ROC curve by calculating

$$E\left(\text{ROC}(t \mid \mathbf{x}) \mid \mathbf{Y}\right) \doteq \frac{1}{m}\sum_{j=1}^{m} \text{ROC}^{(j)}(t \mid \mathbf{x})$$

over a fine grid of values for $t \in (0, 1)$. The covariate-specific AUC and partial AUC are obtained as

$$\text{AUC}(\mathbf{x}) = \int_0^1 \text{ROC}(t \mid \mathbf{x})dt \quad \text{and} \quad \text{pAUC}(\mathbf{x}) = \int_{t_0}^{t_1} \text{ROC}(t \mid \mathbf{x})dt,$$

which are evaluated numerically.

In addition to evaluating test accuracy, the model can be used to calculate predictive risk of disease. Let $\boldsymbol{\theta}$ denote the collection of all model parameters. Risk prediction for an individual with inputs $(\mathbf{x}^*, \mathbf{x}, y)$ is determined by the predictive probability

$$\begin{aligned}
&\Pr(z = 1 \mid \mathbf{x}^*, \mathbf{x}, y, \mathbf{Y}) \\
&= \int \frac{g_1(y - \mathbf{x}'\boldsymbol{\beta})F(\mathbf{x}^{*\prime}\boldsymbol{\alpha})}{g_1(y - \mathbf{x}'\boldsymbol{\beta})F(\mathbf{x}^{*\prime}\boldsymbol{\alpha}) + g_0(y)\{1 - F(\mathbf{x}^{*\prime}\boldsymbol{\alpha})\}} p(d\boldsymbol{\theta} \mid \mathbf{Y}) \\
&\doteq \frac{1}{m}\sum_{j=1}^{m} \frac{g_1^{(j)}(y - \mathbf{x}'\boldsymbol{\beta}^{(j)})F(\mathbf{x}^{*\prime}\boldsymbol{\alpha}^{(j)})}{g_1^{(j)}\left(y - \mathbf{x}'\boldsymbol{\beta}_{(j)}\right)F(\mathbf{x}^{*\prime}\boldsymbol{\alpha}^{(j)}) + g_0^{(j)}(y)\{1 - F(\mathbf{x}^{*\prime}\boldsymbol{\alpha}^{(j)})\}}.
\end{aligned}$$

For fixed $(\mathbf{x}^*, \mathbf{x}, y)$, the odds of disease are given by $g_1(y - \mathbf{x}'\boldsymbol{\beta})F(\mathbf{x}^{*\prime}\boldsymbol{\alpha})[g_0(y)\{1 - F(\mathbf{x}^{*\prime}\boldsymbol{\alpha})\}]^{-1}$, which can be estimated similarly using a numerical approximation of the posterior mean

$$\int \frac{g_1(y - \mathbf{x}'\boldsymbol{\beta})F(\mathbf{x}^{*\prime}\boldsymbol{\alpha})}{g_0(y)\{1 - F(\mathbf{x}^{*\prime}\boldsymbol{\alpha})\}} p(d\boldsymbol{\theta} \mid \mathbf{Y}).$$

Note that for logistic regression, $\exp(\alpha_k)$ is the odds ratio obtained by letting variable k in the vector $\mathbf{x}^*$ increase by one unit, while all other variables (including y and $\mathbf{x}$) are the same for the two groups under comparison, when no interaction is present.

Although not of primary interest, posterior inference for the population prevalence π can be obtained using $\{\mathbf{Z}^{(j)} : j = 1, \ldots, m\}$, where $\mathbf{Z}^{(j)} = (z_1^{(j)}, \ldots, z_n^{(j)})$. Because a random sample of individuals was drawn from the population, $\frac{1}{m}\sum_{j=1}^{m}\left\{\frac{1}{n}\sum_{i=1}^{n} z_i^{(j)}\right\}$ provides an estimate of $\hat{\pi} \equiv \frac{1}{n}\sum_{i=1}^{n} z_i$, the unknown sample proportion of diseased individuals, which is in turn a point estimate of the population prevalence π.

4. Illustrations

For covariate-specific ROC analysis, decisions about which covariates to include in a model are made based on consultations between data analysts and subject-matter experts. This may result in the inclusion of a small number of unnecessary covariates. The simulation in Section 4.2 was therefore conducted to illustrate the robustness of estimates from our model when irrelevant disease and test covariates are included in an analysis. Section 4.3 presents a simulation study of data from nonstandard distributions, and we compare models that have different tree lengths, weight parameters, and centering distributions. We study the use of diffuse priors and increasing sample sizes of simulated data sets in Section 4.4. We begin this section by using our methods to estimate ROC curves and AUCs for the lung cancer study.

4.1. Lung cancer data

We investigated the potential of sEGFR to be a diagnostic biomarker for lung cancer in men. The data were obtained from a case-control study, with 139 cases and 88 controls. To illustrate our models in settings with and without a gold standard, we analyzed the data as if disease status were unknown, using age (the only other variable available to us) as both a disease and test covariate, and we made comparisons with a gold-standard data analysis. Ages (in years) ranged from 35 to 88 for cases and from 24 to 79 for controls. The median (standard deviation) age for controls was 37.5 years (14 years), and for cases, it was 68 years (11 years). Hereafter, age is treated as a continuous variable that has been standardized to have a mean of 0 and a standard deviation of 1. Posterior estimates were based on every 30th iterate from a chain of 100,000 samples, after 5000 draws were discarded as burn-in.

As described in Section 1.1, we found distinct differences in test outcome distributions among controls of different age, but not for the cases. We also found that age may play an important role in helping to discern lung cancer status. We considered a semiparametric model for no-gold-standard data that used age as a test covariate for controls (x = age) and used age as a disease covariate (x^* = age).

We knew beforehand that test accuracy is expected to increase with age and that values of sEGFR tend to be lower for lung cancer cases than controls. Hence, we let $z = 1$ for controls and $z = 0$ for cases. For the no-gold-standard analysis, the sampling models for the natural log-transformed test scores and latent disease status were

$$z_i \sim \text{Bernoulli}(\pi_i), \qquad \text{logit}(\pi_i) = \alpha_0 + \alpha_1 x_i^*,$$
$$f(y_i \mid z_i) = (1 - z_i)g_0(y_i) + z_i g_1(y_i - \beta_0 - \beta_1 x_i).$$

For lung cancer cases, the prior $G_0 \mid (\mu_0, \sigma_0) \sim \text{FPT}_{J_0}(N(\mu_0, \sigma_0^2), c_0)$ was used, with $\mu_0 \sim N(8, 400)$ and $\sigma_0 \sim \text{Uniform}(0, 50)$. The prior mean for μ_0 was chosen because 8 was approximately the sample mean of y for female lung cancer cases in the study, but we allowed for a high degree of prior uncertainty about the value of μ_0 through the large variance of 400. The independent prior on the residual distribution was $G_1 \mid \sigma_1 \sim \text{FPT}_{J_1}(N(0, \sigma_1^2), c_1)$, where G_1 was constrained to have median 0 because otherwise its location would have been confounded with β_0. The prior on σ_1 was Uniform$(0, 50)$, with $\beta_0 \sim N(8, 400)$ and $\beta_1 \sim N(0, 400)$. We compared models with different tree lengths and weight parameters. Specifically, we used $J_0 = J_1$ equal to 4 or 5, and $c_0 = c_1 = 1$ or $c_0, c_1 \sim \text{Gamma}(5, 1)$. Using the same priors, we also considered the underlying parametric normal models. For the parametric analyses, the component distributions in the mixture model were both Gaussian, with $g_0(\cdot)$ being a normal density function with unknown mean μ_0 and variance σ_0^2, and $g_1(\cdot)$ being a mean-zero normal density with unknown variance σ_1^2.

For gold-standard data analysis, the z_i's are known. The data from cases were modeled as independent and identically distributed according to G_0. The regression model for the data from controls was

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, with $\epsilon_i \mid \sigma_1 \sim G_1$. In the semiparametric analysis, the unknown parameters G_0 and G_1 were modeled with the same MFPT priors as in the no-gold-standard data analysis. Similarly, the parametric analysis used the same normal distributions and priors.

Model selection was based on the log pseudo-marginal likelihood (LPML) and corresponding pseudo Bayes factor [48,56]. In the no-gold-standard analysis, the negative LPML (smaller is better) for the parametric normal model was 439, and it was lower for all semiparametric models considered. The negative LPML statistics for the semiparametric models were 422 ($J_0 = J_1 = 4, c_0 = c_1 = 1$), 426 ($J_0 = J_1 = 5, c_0, c_1 \sim \Gamma(5,1)$), 429 ($J_0 = J_1 = 5, c_0 = c_1 = 1$), and 437 ($J_0 = J_1 = 4, c_0, c_1 \sim \text{Gamma}(5,1)$). Compared with the parametric model, the pseudo-Bayes factor of e^{17} decisively supports the selected semiparametric model, which produces data-driven estimates of the component mixture distributions, ROC curves, and AUCs.

Age was found to be a statistically important predictor of disease status, with lung cancer cases tending to be older in age. From the selected semiparametric model applied to no-gold-standard data, the posterior median and 95% PI for α_1 were -1.6 and $(-2.4, -1.0)$, respectively, and $\Pr(\alpha_1 < 0 \mid \mathbf{Y}) \doteq 1$. We also found that the median of the biomarker distribution increases with age and the association between age and sEGFR is statistically important; posterior inferences for β_1 were 0.3 (0.1, 0.6) from the no-gold-standard data analysis, and they were 0.4 (0.2, 0.6) from the gold-standard data analysis.

The accuracy of the biomarker was estimated for 40-, 55-, and 70-year-old men. These ages were the approximate 2.5, 50, and 97.5 empirical percentiles of the age distribution for men in the study. Estimated ROC curves from the gold-standard and no-gold-standard analyses are similar (Figure 1; see also the estimated AUCs in Table I). Discriminatory ability was greater for older men, with semiparametric no-gold-standard estimates of AUC equal to 79%, 83%, and 86% for 40-, 55-, and 70-year-old men, respectively. Differences in AUCs across age were statistically important, since $\Pr(\text{AUC}_{70} > \text{AUC}_j \mid \mathbf{Y}) \doteq 1$ for $j = 40, 55$, and $\Pr(\text{AUC}_{55} > \text{AUC}_{40} \mid \mathbf{Y}) \doteq 1$.

Inferences for age-specific test accuracy based on the parametric normal analysis of no-gold-standard data were different, which underscores the value of a more flexible model for these data; recall that the

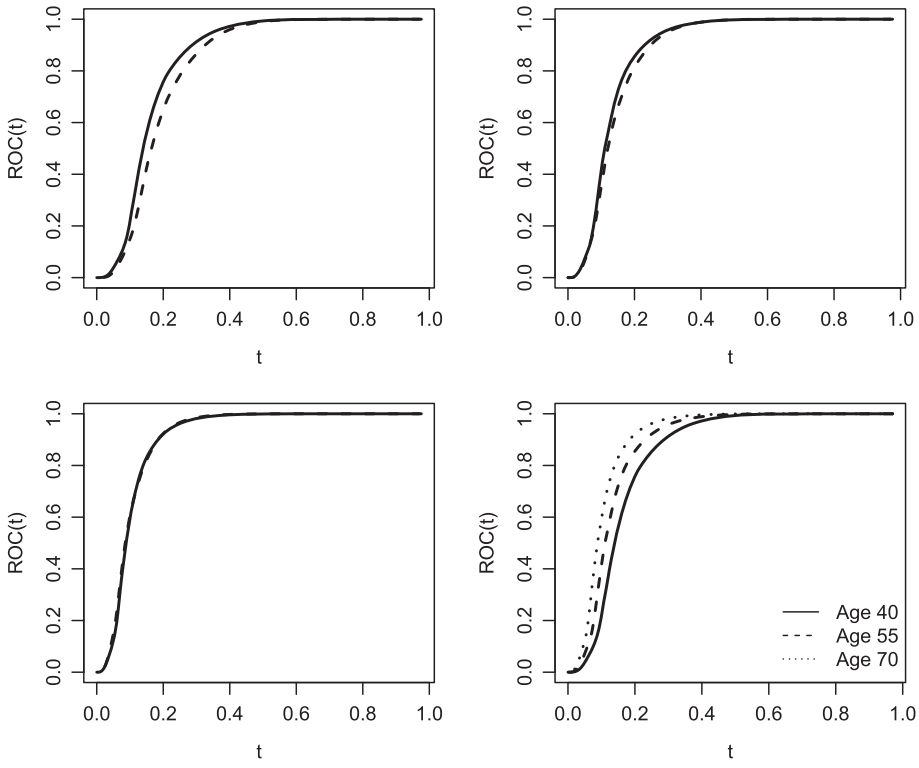

Figure 1. Semiparametric estimates of receiver operating characteristic (ROC) curves for 40- (top-left panel), 55- (top-right panel), and 70-year-old (bottom-left panel) men, when using ln(sEGFR) as a biomarker for lung cancer in a no-gold-standard (solid lines) and gold-standard (dashed lines) data analysis. For ease of comparison across age, the bottom-right panel reproduces the estimated ROC curves from the semiparametric no-gold-standard analysis.

Table I. Semiparametric estimates (posterior median and 95% interval) of AUC and difference in AUC for 40-, 55-, and 70-year-old men, when using ln(sEGFR) as a biomarker of lung cancer in a gold-standard or no-gold-standard data analysis.

Parameter	Gold standard	No gold standard
AUC_{40}	0.78 (0.72, 0.84)	0.79 (0.71, 0.86)
AUC_{55}	0.83 (0.77, 0.88)	0.83 (0.75, 0.89)
AUC_{70}	0.87 (0.81, 0.92)	0.86 (0.77, 0.92)
$AUC_{70} - AUC_{40}$	0.08 (0.05, 0.13)	0.06 (0.02, 0.10)
$AUC_{70} - AUC_{55}$	0.04 (0.02, 0.05)	0.03 (0.01, 0.05)
$AUC_{55} - AUC_{40}$	0.05 (0.03, 0.07)	0.03 (0.01, 0.06)

AUC, area under the curve.

pseudo-Bayes factor strongly supports the Polya tree model. The estimated AUCs from the parametric analyses were 77%, 80%, and 83%, respectively, and all three 95% PIs for pairwise differences between AUCs contained 0. In contrast, none of the three 95% PIs for differences between AUCs contained 0 from the semiparametric analyses.

Our argument for including age as a test covariate for only the non-lung cancer population was based on gold-standard data, which will often be unavailable. In the absence of gold-standard data or strong prior information to support excluding it as a test covariate from the lung cancer population, we would have proceeded by modeling y to depend on age for both the cases and controls. Thus, the previous model of $g_0(y_i)$ for lung cancer cases would be changed to $g_0(y_i - \gamma_0 - \gamma_1 x_i)$. The posterior results support the previous model because the 95% PI for γ_1, $(-0.30, 1.30)$, covers 0. Moreover, the negative LPML for this model is 442, compared with 422 for the previous model (pseudo-Bayes factor $= e^{20}$ in favor of the previous model). Also, estimates of α_1 (posterior median $= -1.7$; 95% PI: $-2.5, -1.1$), β_1 (posterior median $= 0.3$; 95% PI: 0.1, 0.5), and the other model parameters were very similar to estimates from the previous analysis.

4.2. Simulated data: irrelevant covariates

This simulation was set up according to a scenario where the separation between disease and healthy data distributions increases with a covariate, say a_i for age, and the disease probability also increases with a_i. The goal of this simulation study was to investigate if and how estimates change when irrelevant disease and test covariates are included in the model. Three data sets were generated, each with $n = 1000$ where about half the data were from the healthy population and half from the diseased population. In each case, test score data for the healthy population were simulated from the 70–30% mixture of two normal distributions that has density function $0.7(1.5^{-1})\phi(1.5^{-1}(y - 4.5)) + 0.3\phi(y - 5.5)$. The notation $\phi(y)$ indicates the standard normal density function evaluated at y.

The first simulated data set contained five disease covariates, one of which was a_i and the other four were simulated independently from $N(0, 1)$. Latent disease status was simulated independently from Bernoulli(π_i), where $\text{logit}(\pi_i) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i} + \alpha_5 a_i$, for $\boldsymbol{\alpha} = (-5, -3, -2, -4, 2, 1)'$. Test score data for diseased individuals were generated from age-specific normal mixture densities of the form $0.5\phi(y - \beta_0 - \beta_1 a_i, 1) + 0.5\phi(y - \beta_0 - \beta_1 a_i - 4, 1)$, where $\boldsymbol{\beta} = (5, 1)'$ and the test covariates $a_1, \ldots, a_n$ were simulated independently from the Uniform(0, 10) distribution. In case 1, the correct disease and test covariates were included in the model.

Case 2 added two irrelevant disease covariates (x_5 and x_6) to the setting in case 1. The additional covariates were simulated independently from the standard normal distribution, and the true logistic regression coefficient vector was $\boldsymbol{\alpha} = (-5, -3, -2, -4, 2, 1, 0, 0)'$. Case 3 extended the second case by adding an irrelevant binary test covariate (simulated as $b_i \sim \text{Bernoulli}(0.5)$) with true $\boldsymbol{\beta} = (5, 1, 0)'$.

The following inputs were used in all three cases. Tree levels were $J_0 = J_1 = 5$ and $c_0 = c_1 = 1$. The MFPT prior for g_0 was centered at $N(\mu_0, \sigma_0^2)$, while the flexible prior for g_1 was constrained to have median 0 and centered on $N(0, \sigma_1^2)$. A multivariate normal g-prior for $\boldsymbol{\alpha}$ was matched to the beta(8, 8), and the other priors were $\mu_0 \sim \Gamma(2, 0.33)$, $\beta_0 \sim \Gamma(3, 0.33)$, $\beta_1, \beta_2 \sim N(0, 100)$, and $\sigma_0, \sigma_1 \sim \text{Uniform}(0, 20)$. Posterior distributions were approximated using 2000 iterates thinned from 50,000 after a burn-in of 2000.

For $a = b = 0$, Figure 2 shows that pointwise posterior means of densities and ROC curves are very accurate in all three cases. The true AUC is 0.76, and posterior medians and 95% PIs for AUC are 0.74

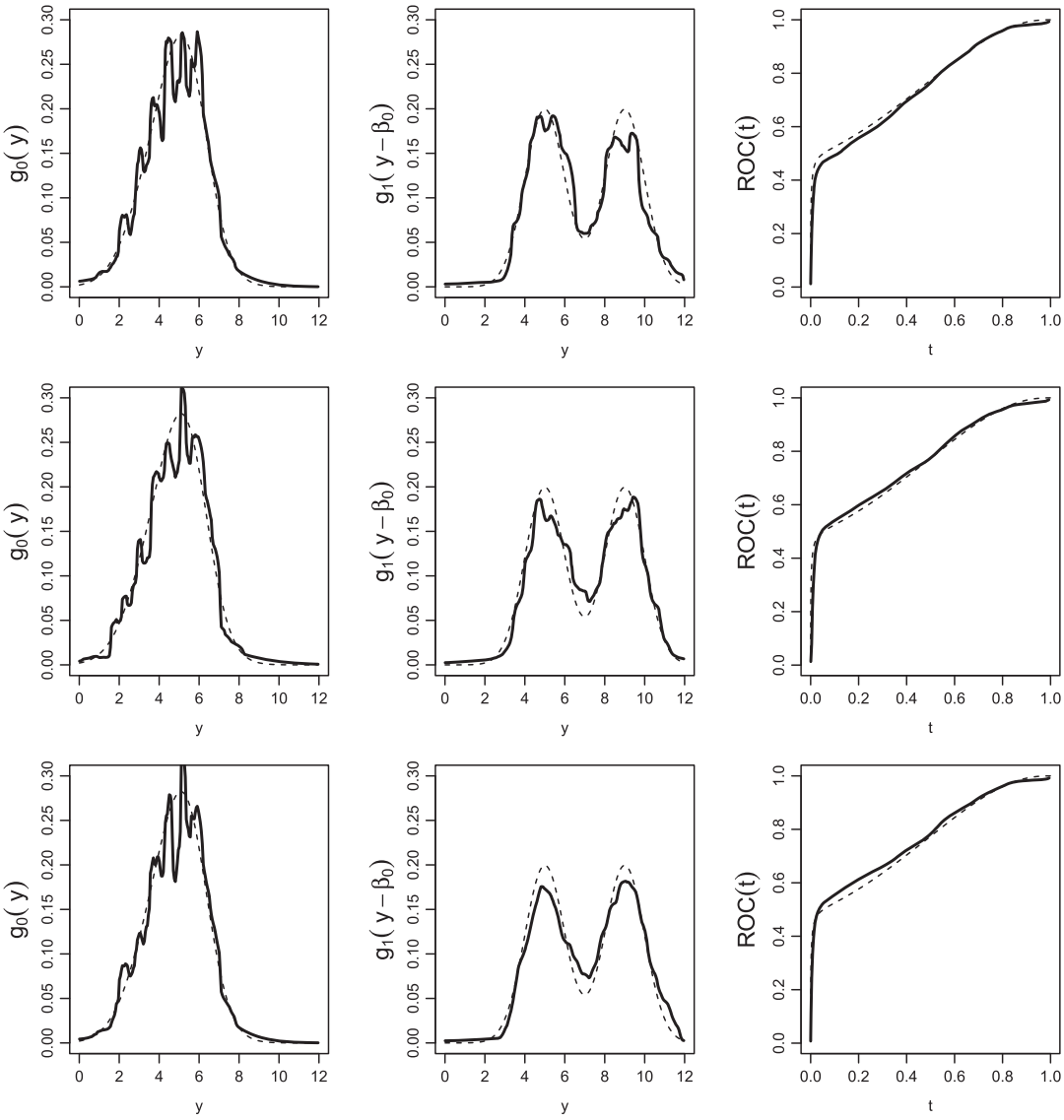

Figure 2. Estimates (solid lines) of density and receiver operating characteristic (ROC) curves when fitting a model that contains (i) the correct five disease covariates and one test covariate (row 1), (ii) seven disease covariates, two of which are insignificant, and the correct test covariate (row 2), and (iii) seven disease covariates, two of which are insignificant, and two test covariates, one of which is insignificant (row 3).

(0.70, 0.78) for case 1, 0.76 (0.71, 0.80) for case 2, and 0.77 (0.69, 0.82) for case 3. Case 2 had true $\alpha_6 = \alpha_7 = 0$, and the posterior means and 95% highest posterior density intervals for α_6 and α_7 were -0.04 $(-0.20, 0.12)$ and -0.05 $(-0.20, 0.12)$, respectively. We note that both of these intervals contain 0, while the 95% highest posterior density intervals for the other five (nonzero) logistic regression coefficients did not contain 0, and those parameters were accurately estimated. Similar conclusions were found for case 3, where the regression coefficients corresponding to the two irrelevant disease covariates and one irrelevant test covariate were estimated to be -0.04 $(-0.21, 0.11)$ for α_6, -0.06 $(-0.22, 0.11)$ for α_7, and -0.01 $(-0.24, 0.28)$ for β_2. The other disease and test covariates in case 3 had regression coefficients that were correctly identified as statistically important, and they were accurately estimated.

4.3. Simulated data: heavy-tail, asymmetric, mixture distributions

In this simulation study, we compared models that contain different tree lengths, weight parameters, and centering families using data ($n = 1000$) generated from distributions that simultaneously have three nonstandard features, namely data from heavy-tailed, skewed mixture distributions. In particular,

g_0 was a two-component 70–30% mixture of noncentral t densities with 10 degrees of freedom; the mixture components had noncentrality parameters of 4.5 and 5.5. Test score data for diseased individuals were generated as in Section 4.2; only noncentral t densities with 10 degrees of freedom were used for mixture components instead of normal densities; that is, data were randomly generated from the mixture density $0.5p_t(y, 10, \beta_0 + \beta_1 a) + 0.5p_t(y, 10, \beta_0 + \beta_1 a + 4)$. Five disease covariates were generated using the same specification as in case 1 in Section 4.2. The priors in the previous section were used here, except β_1 was modeled by a uniform$(0, 20)$ prior (similar results were obtained using lognormal priors that had means of 4.5 or 12 and variances of 35 or 255). Models were fit using at least a 5000 iteration burn-in and between 20,000 to 200,000 post burn-in samples that were thinned by 0, 10, or 50 iterates.

The models under comparison specified equal tree lengths ($J_0 = J_1$) of 4, 5, or 6, with the same normal centering families used in the previous section. The weight parameters were $c_0 = c_1 = 1$ or $c_0, c_1 \sim$ Gamma$(5, 1)$. Negative LPML statistics ranged from a high of 2602 when $J_0 = J_1 = 4$ and $c_0 = c_1 = 1$ to a low of 2560 for the selected model that used $J_0 = J_1 = 5$ and $c_0, c_1 \sim$ Gamma$(5, 1)$. The second lowest negative LPML of 2564 was shared for the model with tree lengths of 4 and gamma priors on the weight parameters and the model with $J_0 = J_1 = 5$ and $c_0 = c_1 = 1$. Pseudo-Bayes factors were >55 in support of the selected model over the other five models.

We also considered models that used t centering families in MFPT priors with $J_0 = J_1 = 4$ and either $c_0 = c_1 = 1$ or $c_0, c_1 \sim$ Gamma$(5, 1)$. The t distributions considered had $\nu_0 = \nu_1 = 3$ degrees of freedom or modeled ν_0 and ν_1 as independent with Gamma$(2, 0.33)$ priors. The negative LPML statistics were larger for these models compared with all six of the models centered at normal families that were considered (the LPML statistics were 2642 and 2645 for the models with fixed and random degrees of freedom, respectively). We note that the models with t centering families required substantially longer computing time compared with the normal-centered models.

Figure 3 presents the accurate estimates of g_0, g_1, and ROC curves from the selected model at test covariates $a = 2, 3$. The AUCs were accurately estimated across a; true AUCs at $a = 0, 1, 2, 3, 4$ are 0.74,

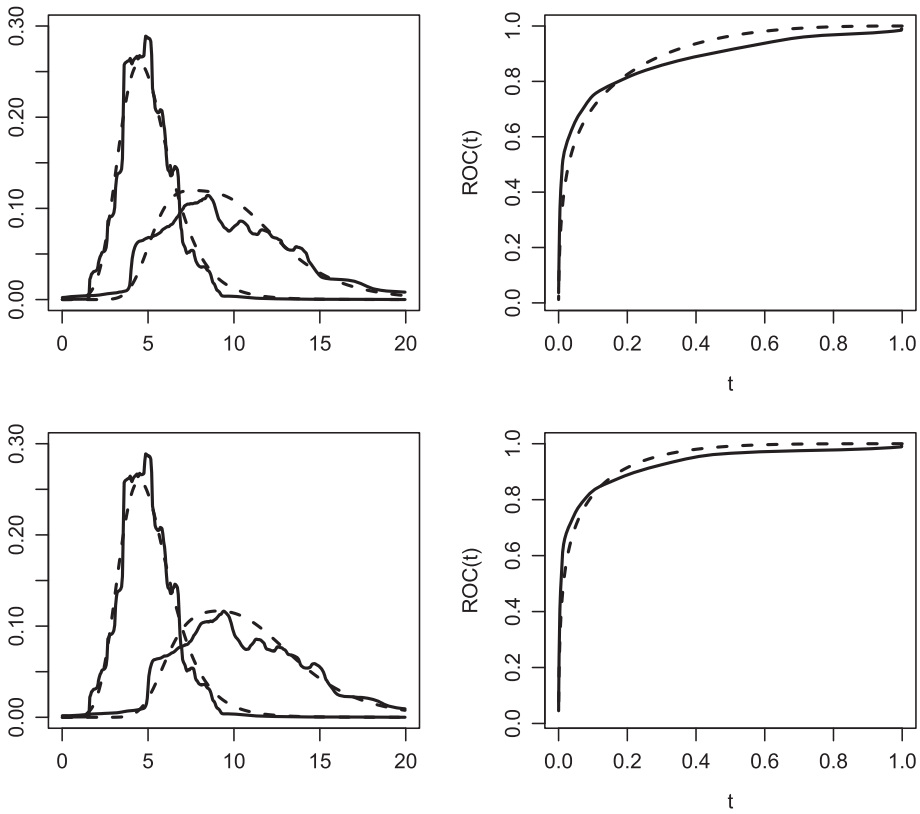

Figure 3. Posterior medians (solid lines) as density and receiver operating characteristic (ROC) curve estimates based on test score data with no gold standard and five disease covariates when the only test covariate is set at $a = 2$ (top row) or $a = 3$ (bottom row). True test-score densities and ROC curves are plotted as dashed lines.

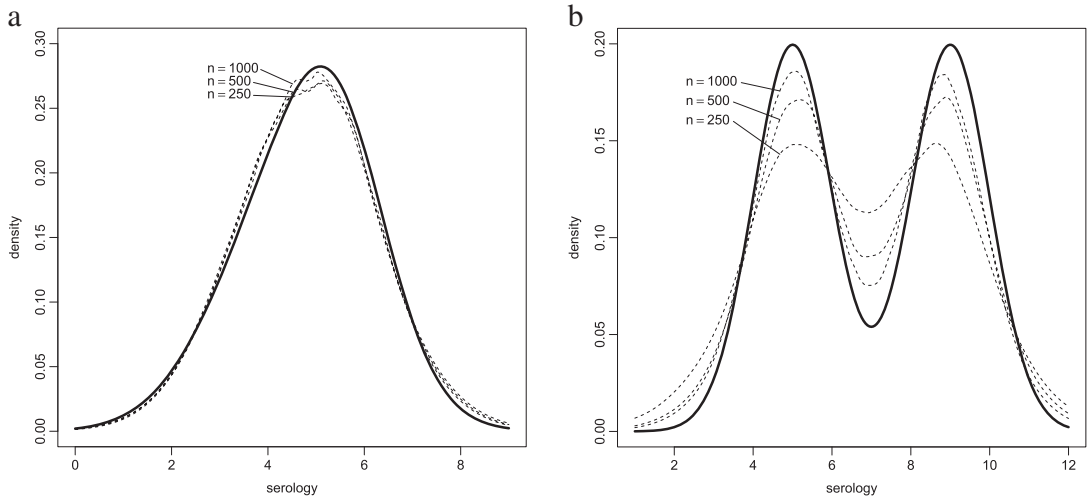

Figure 4. (a) Estimates of $g_0(y-\mu_0)$ (dashed) and truth (solid) for the healthy population. (b) Estimates of $g_1(y-\beta_0)$ (dashed) and truth (solid) for the diseased population.

0.83, 0.90, 0.94, and 0.96, and posterior medians (95% PI) are 0.73 (0.61, 0.79), 0.81 (0.72, 0.86), 0.88 (0.81, 0.93), 0.92 (0.88, 0.95), and 0.95 (0.92, 0.97), respectively.

We personally prefer to use informative priors regardless of whether we have a gold standard or not. However, we note the nice performance when using diffuse priors in the next subsection.

4.4. Simulated data: diffuse priors

Let $\mathbf{x}_i = \mathbf{x}_i^* = (1, a_i)'$, and set $\boldsymbol{\alpha} = (-5, 1)'$ and $\boldsymbol{\beta} = (5, 1)'$. The covariates were generated as $a_i \overset{iid}{\sim} \text{Uniform}(0, 10)$, and the prevalence π was approximately 0.5. Test score data (y) were generated according to the following true densities that were each a mixture of two normals:

$$g_0(y) = 0.7(1.5^{-1})\phi\left((y-4.5)/1.5\right) + 0.3\phi(y-5.5) \quad \text{and} \quad g_1(y) = 0.5\phi(y) + 0.5\phi(y-4).$$

The simulated test scores from the healthy population were slightly skewed, and the diseased population had a test score density with two pronounced modes, perhaps signifying clinical versus subclinical groups, early stage versus latestage disease, or biological subtypes of malignant lung tumors. For each of three samples sizes, $n = 250$, $n = 500$, and $n = 1000$, one hundred data sets were simulated according to the aforementioned specifications, the semiparametric MFPT no-gold-standard model was fit, and posterior densities were approximated, specifically the posterior means of $g_0(y)$ and $g_1(y)$ for 1000 y's over a fine grid.

Figure 4(a) and 4(b) presents the estimated densities averaged over 100 simulated data sets. Clearly, the model is doing an excellent job of estimating g_0 and g_1 as the sample size increases. The estimates appear to be asymptotically unbiased, and although not reported here, all model parameters were estimated correctly, and with increasing precision as n increases. The model appears to estimate g_0 well at any of the three sample sizes, which is not surprising given that the MFPT prior is centered at the normal distribution. The estimates of g_1 improve markedly with increasing sample size. Initially, at $n = 250$, the estimated densities are smoothed more toward the normal centering distribution. As more data are added at $n = 500$ and $n = 1000$, the estimates are able to move closer to the true density (note that roughly half the sample is going into the estimate of g_1, not the whole sample). Note that the bias is greatest at extrema, much like kernel smoothing.

For these simulations, we used truncation levels $J_0 = J_1 = 5$, we modeled $c_0, c_1 \sim \Gamma(5, 1)$, and we used $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_0) \propto 1$ and $p(\sigma_0^2, \sigma_1^2) \propto \sigma_0^{-2}\sigma_1^{-2}$. For each simulated data set, a Gibbs sampler was run 20,000 iterations beyond a burn-in of 1000; every 10-th iterate was thinned for a total of 2000 iterates.

5. Discussion

We developed a general modeling framework for evaluating the performance of continuous medical tests or biomarkers and risk analysis. The models and methods are broadly applicable because they can be

applied to the simple setting involving gold-standard data modeled parametrically without covariates, and to the complicated setting involving a semiparametric regression analysis of no-gold-standard data, and everything in between. Accurate estimates of mixture component distributions from simulated data and ROC curves and AUCs from the lung cancer study were expected given the identifiability of the parametric and semiparametric models for no-gold-standard data.

When gold-standard data are not available, modeling assumptions have to be made to counteract the missing information. It is important to reemphasize the fact that the adequacy of available covariates to reasonably predict latent disease status largely drives the (parametric or semiparametric) model's ability to accurately estimate ROC curves and AUCs. Determination of which covariates to include and how to include them in a model can be aided by consultation with subject-matter experts, but such modeling assumptions can be difficult to verify in this context. Although not considered in this paper because the lung cancer study had only one covariate (age), in situations where many potential covariates are available, it is possible to use methods from Bayesian shrinkage regression (e.g., [57,58]) or a sparseness prior that is a mixture of a parametric distribution and point mass at zero. Clearly, the ideal setting is one in which previous analyses have been performed in the presence of gold-standard data, where it can be established that certain disease covariates would be useful. For example, a previous study used expensive gold-standard data in the development of models for handling future inexpensive no-gold-standard data [59].

We have previously developed an ROC regression model for multivariate gold-standard data using mixtures of multivariate Polya trees [6]. While that model can be extended to handle no-gold-standard data, our application to the lung cancer study involved a univariate response variable, and an extension to multivariate response data is beyond the scope of the current paper. However, this topic is an interesting future line of research.

The utility of semiparametric inference with no-gold-standard data was highlighted in the lung cancer application, where the semiparametric model that was decisively favored by the pseudo-Bayes factor produced different inferences than a parametric normal analysis. Specifically, none of the pairwise differences in AUCs for the three age groups considered were statistically important (based on 95% PIs) according to the parametric analysis, but the AUCs were found to be statistically different based on the semiparametric analysis.

A referee asked us to document the advantages of our MFPT model. The basic idea was to embed traditional parametric models in a broad class of distributions that would allow for departures from the normal–normal model used in the traditional case. The departures allowed are skewness, heavy tails, and multimodality. We expect the same kind of flexibility that one would obtain from a mixture of distributions, only without the necessity to select the number of terms in the mixture. The use of MFPT priors to produce flexible regression models with linear regression structure but much greater flexibility for the error distribution has been well documented over the last decade [46–49]. In addition to modeling error distributions in linear regression models [46], finite or MFPTs have been used to model random effects distributions in linear and generalized linear models [60–62] and to model link functions [47]; they have also been used in nonparametric Rasch models [63] and in modeling multivariate diagnostic outcome data [36], among many other applications. Many of these models have been programmed in the suite of `R` functions, DP Package: Bayesian Nonparametric Modeling in `R`, which can be found at http://cran.r-project.org/web/packages/DPpackage/index.html. Nonparametric frequentist methods for gold-standard ROC analysis using location-scale models are also available [64].

Appendix A: Identifiability

We provide a heuristic argument for identifiability of our semiparametric model. To accomplish this, we first begin with its parametric counterpart, which has been asserted in the literature to be identifiable (identifiability of the normal version is proven in [50]). Then we extend the argument to the semiparametric case.

A.1 Parametric case

Let $(\mathbf{Y}, \mathbf{X}, \mathbf{X}^*, \mathbf{z})$ denote the augmented data, where $\mathbf{Y}$ is an $n \times 1$ vector of observed test scores, $\mathbf{X}$ is an $n \times (p+1)$ matrix of test covariate values, $\mathbf{X}^*$ is an $n \times (s+1)$ matrix of disease covariate values, and $\mathbf{z}$ denotes the latent vector that provides disease status for each individual in the sample. Matrices $\mathbf{X}$ and

$\mathbf{X}^*$ contain a first column of all ones to accommodate intercepts. We assume that the (column) ranks of $\mathbf{X}$ and $\mathbf{X}^*$ are full, and that individuals are sampled independently. Furthermore, we assume a generic model for the moment with

$$g(\mathbf{Y}, \mathbf{z} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} g(y_i, z_i \mid \mathbf{x}_i, \mathbf{x}_i^*, \boldsymbol{\theta}) = \prod_{i=1}^{n} g(y_i \mid \mathbf{x}_i, z_i, \boldsymbol{\theta}) g(z_i \mid \mathbf{x}_i^*, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is a parameter vector that encompasses the entire model. We assume

$$z_i \sim \text{Bernoulli}(F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha})),$$

where $F(\cdot)$ is a CDF and is usually selected as the standard logistic, which results in a logistic regression model for latent disease status. We also define

$$g(y_i \mid \mathbf{x}_i, z_i = 0, \boldsymbol{\theta}) = g_0(y_i - \mu \mid \boldsymbol{\lambda}_0) \text{ and } g(y_i \mid \mathbf{x}_i, z_i = 1, \boldsymbol{\theta}) = g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta} \mid \boldsymbol{\lambda}_1),$$

where $\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1)$. The parameter vectors $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ are regarded as corresponding to, for example, scale and skewness. For example, with both g_0 and g_1 as normal densities, $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_1$ are variance parameters. Thus, g_0 and g_1 are not necessarily in the same family, but they are parametric location family densities. Moreover, the model corresponding to g_0 is free of covariates, and the model corresponding to g_1 depends on $\mathbf{x}_i'\boldsymbol{\beta}$, which is a location parameter on the scale of the data (or transformed data) y_i. The marginal model for $y_i \mid \mathbf{x}_i^*, \mathbf{x}_i, \boldsymbol{\theta}$ is thus

$$g(y_i \mid \mathbf{x}_i^*, \mathbf{x}_i, \boldsymbol{\theta}) = F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}) g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta} \mid \boldsymbol{\lambda}_1) + \{1 - F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha})\} g_0(y_i - \mu \mid \boldsymbol{\lambda}_0).$$

We assume that the marginal models, $g_0(\cdot \mid \boldsymbol{\lambda}_0)$ and $g_1(\cdot \mid \boldsymbol{\lambda}_1)$, are identifiable. Specifically, with given $(\boldsymbol{\lambda}_{01}, \boldsymbol{\lambda}_{11})$ and $(\boldsymbol{\lambda}_{02}, \boldsymbol{\lambda}_{12})$, if we have $g_j(u \mid \boldsymbol{\lambda}_{j1}) = g_j(u \mid \boldsymbol{\lambda}_{j2})$ for all possible scalars u, then we must have $(\boldsymbol{\lambda}_{01}, \boldsymbol{\lambda}_{11}) = (\boldsymbol{\lambda}_{02}, \boldsymbol{\lambda}_{12})$, for $j = 0, 1$.

Now consider two parameter vectors, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, where $\boldsymbol{\theta}_k = (\mu_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \boldsymbol{\lambda}_{0k}, \boldsymbol{\lambda}_{1k})$ for $k = 1, 2$. Assume that

$$g(\mathbf{Y} \mid \mathbf{x}, \mathbf{x}^*, \boldsymbol{\theta}_1) = g(\mathbf{Y} \mid \mathbf{x}, \mathbf{x}^*, \boldsymbol{\theta}_2), \tag{A.1}$$

for all possible vectors $\mathbf{Y}$. Our goal is to find conditions under which this cannot happen unless $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. If $F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}) = \pi$ for all i, and if the only distinction between g_0 and g_1 is an unknown location, it is well known that the aforementioned mixture model lacks identifiability. The model is identifiable if it is constrained to have one of the locations smaller than the other. This is the classic 'label-switching' problem. It is also known that if there are no covariates, and if g_0 and g_1 are modeled nonparametrically, then the corresponding mixture model lacks identifiability, even if one distribution stochastically dominates the other. We proceed to argue that the aforementioned model with dependence on covariates is identifiable under some mild conditions.

Under our assumptions, it follows that (A.1) holds if and only if

$$\begin{aligned}&\textstyle\prod_{i=1}^{n} \left[\{1 - F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_1)\} g_0(y_i - \mu_1 \mid \boldsymbol{\theta}_1) + F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_1) g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}_1 \mid \boldsymbol{\theta}_1)\right] \\ &= \textstyle\prod_{i=1}^{n} \left[\{1 - F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_2)\} g_0(y_i - \mu_2 \mid \boldsymbol{\theta}_2) + F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_2) g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}_2 \mid \boldsymbol{\theta}_2)\right],\end{aligned}$$

which holds if and only if

$$\begin{aligned}&\textstyle\sum_{\mathbf{Z} \in \{0,1\}^n} \left[\prod_{i=1}^{n} \left\{F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_1)^{z_i} (1 - F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_1))^{1-z_i} g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}_1 \mid \boldsymbol{\theta}_1)^{z_i} g_0(y_i - \mu_1 \mid \boldsymbol{\theta}_1)^{1-z_i}\right\}\right] \\ &= \textstyle\sum_{\mathbf{Z} \in \{0,1\}^n} \left[\prod_{i=1}^{n} \left\{F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_2)^{z_i} (1 - F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_2))^{1-z_i} g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}_2 \mid \boldsymbol{\theta}_2)^{z_i} g_0(y_i - \mu_2 \mid \boldsymbol{\theta}_2)^{1-z_i}\right\}\right].\end{aligned}$$

This equation will hold in two cases, namely if and only if, for all i,

$$F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_1) = F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_2),\ g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}_1 \mid \boldsymbol{\lambda}_{11}) = g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}_2 \mid \boldsymbol{\lambda}_{12}), g_0(y_i - \mu_1 \mid \boldsymbol{\lambda}_{01}) = g_0(y_i - \mu_2 \mid \boldsymbol{\lambda}_{02}), \tag{A.2}$$

or

$$F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_1) = 1 - F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_2),\ g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}_1 \mid \lambda_{11}) = g_0(y_i - \mu_2 \mid \lambda_{02}),\ g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}_2 \mid \lambda_{12}) = g_0(y_i - \mu_1 \mid \lambda_{01}). \tag{A.3}$$

Case (A.2) can only occur if $\mathbf{x}^{*\prime}\boldsymbol{\alpha}_1 = \mathbf{x}^{*\prime}\boldsymbol{\alpha}_2$, and if $\lambda_{j1} = \lambda_{j2}$ for $j = 0, 1$, $\mathbf{x}'\boldsymbol{\beta}_1 = \mathbf{x}'\boldsymbol{\beta}_2$, and $\mu_1 = \mu_2$. Therefore, case (A.2) requires that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

Case (A.3) cannot hold unless $g_0 = g_1$ and $\lambda_{j1} = \lambda_{j2}$ for $j = 0, 1$. However, if $g_0 = g_1$ and $\lambda_1 = \lambda_2$, then the equalities in (A.3) can only occur if $\mathbf{x}^{*\prime}\boldsymbol{\alpha}_1 = -\mathbf{x}^{*\prime}\boldsymbol{\alpha}_2$, for symmetric F, and $\mathbf{x}'\boldsymbol{\beta}_1 = \mathbf{x}'\boldsymbol{\beta}_2 = \mu_1 = \mu_2$. Then we must have $\boldsymbol{\alpha}_1 = -\boldsymbol{\alpha}_2$, for symmetric F, and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, because $\mathbf{x}^*$ and $\mathbf{x}$ have full column rank. If the rank and dimension of both $\mathbf{x}^*$ and $\mathbf{x}$ are one, then the model clearly lacks identifiability since we are in the previously described label-switching situation. We make the label switching impossible by forcing $\mu \leqslant \beta_0$, where β_0 is the intercept (and location parameter) for the g_1 population. Under this constraint, (A.3) is vacuous, and thus, the model is identifiable.

When the rank of $\mathbf{x}^*$ is greater than 1, we can restrict the model in such a way that it is impossible for $\boldsymbol{\alpha}_2 = -\boldsymbol{\alpha}_1$, in which case (A.3) is again vacuous and the model will be identifiable. We could merely constrain one of the coefficients to be positive based on subjective considerations. This kind of lack of identifiability is quite mild and would rarely show up in applications as problematic even without any restrictions, provided informative priors were used for the disease covariate regression coefficients. We also note that this result implies there would be two sets of maximum likelihood estimates (MLEs). For example, if $(\hat{\mu}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}^*)$ is an MLE for $\boldsymbol{\theta}$, then so is $(\hat{\mu}, -\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}^*)$. Thus, if one obtained an MLE for $\boldsymbol{\alpha}$ that made no sense while the negative of it did, one would simply use the version that did make sense.

Concluding results: (i) If $g_0 = g_1$ and the rank of $\mathbf{x}^*$ is greater than 1, with symmetric F, and with a constraint on $\boldsymbol{\alpha}$ that eliminates the possibility that $\boldsymbol{\alpha}_2 = -\boldsymbol{\alpha}_1$, the parametric mixture model is identifiable. (ii) If $g_0 \neq g_1$, the model is automatically identifiable because (A.3) is void in this case and the only way that (A.2) can hold is if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. For instance, if we model g_0 as normal and g_1 as Weibull, the model will be identifiable without any restrictions. (iii) If F is not symmetric and the rank of $\mathbf{x}^*$ is greater than one, the model is identifiable.

A.2 Semiparametric case

We consider the same situation as in the preceding text, only now we replace λ_j with $\chi_j = (\boldsymbol{\Pi}_j, \lambda_j)$ for $j = 0, 1$, where $\boldsymbol{\Pi}_j$ is the set of branch probabilities and λ_j now is a vector of parameters in the centering parametric family of the Polya tree for population j. For fixed c, the density g_0 is completely determined by χ_0, and similarly, the density g_1 is completely determined by χ_1. While $\boldsymbol{\Pi}_0$ and $\boldsymbol{\Pi}_1$ are theoretically infinite dimensional, we only use a finite dimensional version, but the dimension of $\boldsymbol{\Pi}_j$ is meant to be quite large compared with the dimension of λ_j. Recall that χ_1 is defined so that g_1 has median zero; there is no such constraint on χ_0.

Now consider two parameter vectors, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, where $\boldsymbol{\theta}_k = (\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \chi_{0k}, \chi_{1k})$, for $k = 1, 2$. Assume that (A.1) holds for these values. Then, as in the development in the preceding text, this will hold if and only if, for all i,

$$\begin{gathered} F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_1) = F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_2) \\ g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}_1 \mid \chi_{11}) = g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}_2 \mid \chi_{12}), \quad g_0(y_i \mid \chi_{01}) = g_0(y_i \mid \chi_{02}), \end{gathered} \tag{A.4}$$

or

$$\begin{gathered} F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_1) = 1 - F(\mathbf{x}_i^{*\prime}\boldsymbol{\alpha}_2) \\ g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}_1 \mid \chi_{11}) = g_0(y_i \mid \chi_{02}),\ g_1(y_i - \mathbf{x}_i'\boldsymbol{\beta}_2 \mid \chi_{12}) = g_0(y_i \mid \chi_{01}). \end{gathered} \tag{A.5}$$

For (A.4) or (A.5) to hold, we must have $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2$ in the former case, or $\boldsymbol{\alpha}_1 = -\boldsymbol{\alpha}_2$ in the latter case when F is symmetric. But if F is not symmetric, or if we place a restriction on $\boldsymbol{\alpha}$ as discussed in the parametric case, then (A.5) is void, and we are only concerned with (A.4). But the only way (A.4) can hold is if $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2$, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, $\chi_{11} = \chi_{12}$, and $\chi_{01} = \chi_{02}$.

Concluding results: (i) If the rank of $\mathbf{x}^*$ is greater than 1, and with a constraint on $\boldsymbol{\alpha}$ that eliminates the possibility that $\boldsymbol{\alpha}_1 = -\boldsymbol{\alpha}_2$, the semiparametric mixture model is identifiable. Note that the model is

still identifiable even if the rank of $\mathbf{x}$ is one, so there need not be any test covariates. This implies that the model in [4] is identifiable under a constraint on $\boldsymbol{\alpha}$. (ii) Now suppose the model $g_0(y_i)$ is replaced with $g_0(y_i - \tilde{\mathbf{x}}_i'\boldsymbol{\gamma} \mid \chi_0)$, that is, suppose there are test covariates that also affect the distribution of outcomes in the absence of disease. Augment χ_k with $\boldsymbol{\gamma}_k$, $k = 1, 2$. Then (A.4) and (A.5) can be suitably modified. We assume in this instance that g_0 has median zero. Condition (A.5) does not exist with suitable constraint on $\boldsymbol{\alpha}$. Condition (A.4) is similar as before except now we add that $\tilde{\mathbf{x}}\boldsymbol{\gamma}_1 = \tilde{\mathbf{x}}\boldsymbol{\gamma}_2$, which occurs only if $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2$. So adding test covariates leaves identifiability alone, under constraints on $\boldsymbol{\alpha}$.

Appendix B: SAS code

SAS 9.3 template code to fit the ROC regression model is presented in the succeeding text. As an example, the code is presented for settings with two disease covariates (d_1 and d_2) and one test covariate (t); simple modifications are needed for other settings. In the SAS data step (not shown), variables `y`, `d1`, `d2`, and `t` are read into a data set that was named `ROC`. The `iml` procedure is used to calculate vector $\boldsymbol{be}$ and matrix $\boldsymbol{\Sigma} = gn(\mathbf{X}^{*\prime}\mathbf{X}^{*})^{-1}$ in the $N_3(\boldsymbol{be}, \boldsymbol{\Sigma})$ g-prior on $\boldsymbol{\alpha}$ that matches a beta(a_π, b_π) distribution of the users choosing (the user inputs values for `api` and `bpi`). The first line of the `mcmc` procedure instructs SAS to use the data in `ROC` and specifies options for the number of burn-in iterates (`nbi`), the number of iterates to simulate post burn-in (`nmc`), the amount to thin (`thin`), and the parameters to monitor for posterior inference. The next segment of code defines constants that are used in the program. For illustration, tree levels were set to 3, and weight parameters were set to 1. The branch probability `q11` was set to 0.5 to constrain the median of G_1 to be 0. The user inputs the values for `be` and `a,b,...,i` that were output by `proc iml`. Branch probabilities for the healthy and diseased groups are coded by `pjk` and `qjk`, with the appropriate numbers substituted for `j` and `k`. The marginal probabilities at level 3 are named `p0k` and `q0k`, where `k` ranges from 1 to 8. Priors on branch probabilities, regression coefficients, and parameters of the centering distributions follow. The log likelihood (`llike`) involves `k0` and `k1`, which identify the sets each y_i belongs to at level 3 of the trees. This code gives the log of the density $f_i(y) = (1 - \pi_i)g_0(y) + \pi_i g_1(y - \beta_0 - \beta_1 t_i)$. The model is fit using Gibbs sampling with block updating occurring for the groups of parameters that are defined in the `parms` statements.

```
proc iml;
 use ROC;
 read all var{d1 d2} into Xtemp;
 n=nrow(Xtemp);
 X=(j(n,1,1))||Xtemp; * Adds to Xtemp an nx1 vector of all ones;
 p=ncol(X);
 api=; bpi=;
 b=digamma(api)-digamma(bpi);
 g=(trigamma(api)+trigamma(bpi))/p;
 S=inv(X`*X);
 be=j(p,1,0); be[1,1]=b;
 Sigma=S#(n*g);
 print be Sigma;
quit;

proc mcmc data=ROC nbi= nmc= thin= propcov= outpost=out
          monitor=(p0 q0 mu0 sigma0 beta0 beta1 sigma1 alpha1 alpha2
          alpha3);
begincnst;
J0=3; J1=3; c0=1; c1=1; q11=0.5;
array be[3] (0 0 0);
array S[3,3] (
a b c
d e f
g h i
)
;
```

```
endcnst;
array alpha[3];
parms p11 p21 p23 p31 p33 p35 p37;
parms q21 q23 q31 q33 q35 q37;
parms mu0 sigma0;
parms beta0 beta1 sigma1;
parms alpha;

array p0[8];
p01=p11*p21*p31;          p02=p11*p21*(1-p31);
p03=p11*(1-p21)*p33;      p04=p11*(1-p21)*(1-p33);
p05=(1-p11)*p23*p35;      p06=(1-p11)*p23*(1-p35);
p07=(1-p11)*(1-p23)*p37;  p08=(1-p11)*(1-p23)*(1-p37);

array q0[8];
q01=q11*q21*q31;          q02=q11*q21*(1-q31);
q03=q11*(1-q21)*q33;      q04=q11*(1-q21)*(1-q33);
q05=(1-q11)*q23*q35;      q06=(1-q11)*q23*(1-q35);
q07=(1-q11)*(1-q23)*q37;  q08=(1-q11)*(1-q23)*(1-q37);

prior p11 ~ beta(c0*1**2, c0*1**2);
prior p21 p23 ~ beta(c0*2**2, c0*2**2);
prior p31 p33 p35 p37 ~ beta(c0*3**2, c0*3**2);
prior q21 q23 ~ beta(c1*2**2, c1*2**2);
prior q31 q33 q35 q37 ~ beta(c1*3**2, c1*3**2);
prior mu0 ~ ; prior beta0 ~ ; prior beta1~ ; prior sigma0 ~ ;
prior sigma1 ; prior alpha ~ mvn(be, S);

pi = exp(alpha1+alpha2*d1+alpha3*d2)/(1+exp(alpha1+alpha2*d1
     +alpha3*d2));
k0 = int(2**J0 * cdf("normal", y, mu0, sigma0) + 1);
k1 = int(2**J1 * cdf("normal", y-(beta0+beta1*t), 0, sigma1) + 1);
llike = log((1-pi)*(2**J0)*(p0[k0])*pdf("normal",y,mu0,sigma0)
        + pi*(2**J1)*(q0[k1])*pdf("normal",y-(beta0+beta1*t),0,
        sigma1));
model general(llike);
run;
```

References

1. Erkanli A, Sung M, Costello EJ, Angold A. Bayesian semi-parametric ROC analysis. *Statistics in Medicine* 2006; **25**: 3905–3928.
2. Wang C, Turnbull BW, Gröhn YT, Nielsen SS. Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown. *Journal of Agricultural, Biological, and Environmental Statistics* 2007; **12**:128–146.
3. Fosgate GT, Scott HM, Jordan ER. Development of a method for Bayesian nonparametric ROC analysis with application to an ELISA for Johne's disease in dairy cattle. *Preventive Veterinary Medicine* 2007; **81**:178–193.
4. Branscum AJ, Johnson WO, Hanson TE, Gardner IA. Bayesian semiparametric ROC curve estimation and disease diagnosis. *Statistics in Medicine* 2008; **27**:2474–2496.
5. Hanson TE, Kottas A, Branscum AJ. Modelling stochastic order in the analysis of ROC data: Bayesian non-parametric approaches. *Applied Statistics* 2008; **57**:207–225.
6. Hanson TE, Branscum AJ, Gardner IA. Multivariate mixtures of Polya trees for modelling ROC data. *Statistical Modelling* 2008; **8**:81–96.
7. Gu J, Ghosal S, Roy A. Bayesian bootstrap estimation of ROC curve. *Statistics in Medicine* 2008; **27**:5407–5420.
8. Inácio V, Turkman AA, Nakas CT, Alonzo TA. Nonparametric Bayesian estimation of the three-way receiver operating characteristic surface. *Biometrical Journal* 2011; **53**:1011–1024.
9. Ladouceur L, Rahme E, Belisle P, Scott AN, Schwartzman K, Joseph L. Modeling continuous diagnostic test data using approximate Dirichlet process distributions. *Statistics in Medicine* 2011; **30**:2648–2662.
10. Hung H, Chiang C-T. Nonparametric methodology for the time-dependent partial area under the ROC curve. *Journal of Statistical Planning and Inference* 2011; **141**:3829–3838.

11. Colak E, Mutlu F, Bal C, Oner S, Ozdamar K, Gok B, Cavusoglu Y. Comparison of semiparametric, parametric, and non-parametric ROC analysis for continuous diagnostic tests using a simulation study and acute coronary syndrome data. *Computational and Mathematical Methods in Medicine* 2012; **2012**:Article ID 698320, 7 pages. DOI:10.1155/2012/698320.
12. Hanfang Y, Zhao Y. Smoothed empirical likelihood for ROC curves with censored data. *Journal of Multivariate Analysis* 2012; **109**:254–263.
13. Inácio de Carvalho V, Jara A, Hanson TE, de Carvalho M. Bayesian nonparametric ROC regression modeling. *Bayesian Analysis* 2013; **8**:623–646.
14. Martinez-Camblora P, Carleos C, Corral N. General nonparametric ROC curve comparison. *Journal of the Korean Statistical Society* 2013; **43**:71–81.
15. Broemeling LD. *Bayesian Biostatistics and Diagnostic Medicine*. Chapman & Hall: Boca Raton, FL, 2007.
16. Collins J, Huynh M. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Statistics in Medicine* 2014; **33**:4141–4169.
17. Gastwirth JL, Johnson WO. Quality control for screening tests: applications to HIV and drug use detection. *Journal of the American Statistical Association* 1994; **89**:972–981.
18. Johnson WO, Gastwirth JL. Dual group screening. *Journal of Statistical Planning and Inference* 2000; **83**:449–473.
19. Baron AT, Lafky JM, Boardman CH, Balasubramaniam S, Suman VJ, Podratz KC, Maihle NJ. Serum sErbB1 and epidermal growth factor levels as tumor biomarkers in women with stage III or IV epithelial ovarian cancer. *Cancer Epidemiology Biomarkers and Prevention* 1999; **8**:129–137.
20. Baron AT, Boardman CH, Lafky JM, Rademaker A, Liu D, Fishman DA, Podratz KC, Maihle NJ. Soluble epidermal growth factor receptor (sEGFR) [corrected] and cancer antigen 125 (CA125) as screening and diagnostic tests for epithelial ovarian cancer. *Cancer Epidemiology Biomarkers and Prevention* 2005; **14**:306–318.
21. Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 2004; **60**:427–435.
22. Jones G, Johnson WO, Hanson TE, Christensen R. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics* 2010; **66**:855–863.
23. Box GEP. Science and statistics. *Journal of the American Statistical Association* 1976; **71**:791–799.
24. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford Statistical Science Series. Oxford University Press, 2003.
25. Krzanowski WJ, Hand DJ. *ROC Curves for Continuous Data*. Chapman & Hall: Boca Raton, FL, 2009.
26. Henkelman RM, Kay I, Bronskill MJ. Receiver operating characteristic (ROC) without truth. *Medical Decision Making* 1990; **10**:24–29.
27. Beiden SV, Campbell G, Meier K, Wagner RF. On the problem of ROC analysis without truth: the EM algorithm and the information matrix. In *Medical Imaging 2000: Image Perception and Performance*, Krupinski EA (ed.) SPIE Digital Library: San Diego, CA, 2000; 126–134.
28. Kupinski MA, Hoppin JW, Clarkson E, Barrett HH, Kastis GA. Estimation in medical imaging without a gold standard. *Academic Radiology* 2002; **9**:290–297.
29. Choi YK, Johnson WO, Collins MT, Gardner IA. Bayesian estimation of ROC curves in the absence of a gold standard. *Journal of Agricultural, Biological, and Environmental Statistics* 2006; **11**:210–229.
30. Choi YK, Johnson WO, Thurmond MC. Diagnosis using predictive probabilities without cut-offs. *Statistics in Medicine* 2006; **25**:699–717.
31. Hall P, Zhou X-H. Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics* 2003; **31**:201–224.
32. Zhou X-H, Castelluccio P, Zhou C. Nonparametric estimation of ROC curves in the absence of a gold standard. *Biometrics* 2005; **61**:600–609.
33. Albert PS. Random effects modeling approaches for estimating ROC curves from repeated ordinal tests without a gold standard. *Biometrics* 2007; **63**:593–602.
34. Wang C, Turnbull BW, Gröhn YT, Nielsen SS. Estimating receiver operating characteristic curves with covariates when there is no perfect reference test for diagnosis of Johne's disease. *Journal of Dairy Science* 2006; **89**:3038–3046.
35. Wang C, Turnbull BW, Nielsen SS, Gröhn YT. Bayesian analysis of longitudinal Johne's disease diagnostic data without a gold standard test. *Journal of Dairy Science* 2006; **94**:2320–2328.
36. Norris M, Johnson WO, Gardner IA. Modeling bivariate longitudinal diagnostic outcome data in the absence of a gold standard. *Statistics and Its Interface* 2009; **2**:171–185.
37. Branscum AJ, Johnson WJ, Baron AT. Robust medical test evaluation using flexible Bayesian semiparametric regression models. *Epidemiology Research International* 2013; **2013**:Article ID 131232.
38. Rodriguez A, Martinez JC. Bayesian semiparametric estimation of covariate-dependent ROC curves. *Biostatistics* 2014; **2**:353–369.
39. Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* 1998; **54**:124–135.
40. Rodríguez-Álvarez MX, Tahoces PG, Cadarso-Suárez C, Lado MJ. Comparative study of ROC regression techniques – applications for the computer-aided diagnostic system in breast cancer detection. *Computational Statistics and Data Analysis* 2011; **55**:888–902.
41. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology* 1997; **146**:195–203.
42. McInturff P, Johnson WO, Gardner IA, Cowling DW. Bayesian modeling of risk based on outcomes that are subject to error. *Statistics in Medicine* 2004; **23**:1095–1107.
43. Ferguson TS. Prior distributions on spaces of probability measures. *Annals of Statistics* 1974; **2**:615–629.
44. Lavine M. Some aspects of Polya tree distributions for statistical modeling. *Annals of Statistics* 1992; **20**:1222–1235.
45. Lavine M. More aspects of Polya tree distributions for statistical modeling. *Annals of Statistics* 1994; **22**:1161–1176.

46. Hanson TE, Johnson WO. Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association* 2002; **97**:1020–1033.
47. Hanson TE. Inferences for mixtures of finite Polya trees. *Journal of the American Statistical Association* 2006; **101**: 1548–1565.
48. Christensen R, Johnson W, Branscum A, Hanson TE. *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. CRC Press: Boca Raton, FL, 2010.
49. Christensen R, Hanson T, Jara A. Parametric nonparametric statistics: an introduction to mixtures of finite Polya trees. *The American Statistician* 2008; **62**:296–306.
50. Jiang W, Tanner MA. On the identifiability of mixtures-of-experts. *Neural Networks* 1999; **12**:197–220.
51. Faraggi D. Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician* 2003; **52**:179–192.
52. Bedrick EJ, Christensen R, Johnson WO. A new perspective on priors for generalized linear models. *Journal of the American Statistical Association* 1996; **91**:1450–1460.
53. Zhang B, Chen Z, Albert PS. Estimating diagnostic accuracy of raters without a gold standard by exploiting a group of experts. *Biometrics* 2012; **68**:1294–1302.
54. Hanson TE, Branscum AJ, Johnson WJ. Informative *g*-priors for logistic regression. *Bayesian Analysis* 2014; **9**:597–612.
55. Gelman A, Jakulin A, Pittau MG, Su Y-S. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics* 2008; **2**:1360–1383.
56. Geisser S, Eddy W. A predictive approach to model selection. *Journal of the American Statistical Association* 1979; **74**:153–60.
57. Park T, Casella G. The Bayesian lasso. *Journal of the American Statistical Association* 2008; **103**:681–686.
58. Huang A, Xu S, Cai X. Empirical Bayesian LASSO-logistic regression for multiple binary trait locus mapping. *BMC Genetics* 2013; **14**:5.
59. Thurmond MC, Johnson WO, Muñoz-Zanzi C, Su CL, Hietala S. Probability diagnostic assignment for serologic measures, with application to *Neospora caninum* infection. *American Journal of Veterinary Research* 2002; **63**:318–325.
60. Walker SG, Mallick BK. Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society, Series B* 1997; **59**:845–860.
61. Branscum AJ, Hanson TE. Bayesian nonparametric meta-analysis using Polya tree mixture models. *Biometrics* 2008; **64**:825–833.
62. Jara A, Hanson TE, Lesaffre E. Robustifying generalized linear mixed models using a new class of mixtures of multivariate Polya trees. *Journal of Computational and Graphical Statistics* 2009; **18**:838–860.
63. San Martin E, Jara A, Rolin J-M, Mouchart M. On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika* 2011; **76**:385–409.
64. Gonzalez-Manteiga W, Pardo-Fernandez JC, Van Keilegom I. ROC curves in non-parametric location-scale regression models. *Scandinavian Journal of Statistics* 2011; **38**:169–184.