Published in final edited form as:

Biometrics. 2016 June ; 72(2): 554–562. doi:10.1111/biom.12453.

A unified Bayesian semiparametric approach to assess discrimination ability in survival analysis

Lili Zhao,

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A

Dai Feng,

Biometrics Research Department, Merck Research Laboratories, U.S.A

Guoan Chen, and

Section of Thoracic Surgery, Department of Surgery, University of Michigan, Ann Arbor, Michigan, U.S.A

Jeremy M.G. Taylor

Department of Biostatistics, University of Michigan, Ann Arbor

Lili Zhao: zhaolili@umich.edu

Summary

The discriminatory ability of a marker for censored survival data is routinely assessed by the time-dependent ROC curve and the c -index. The time-dependent ROC curve evaluates the ability of a biomarker to predict whether a patient lives past a particular time t . The c -index measures the global concordance of the marker and the survival time regardless of the time point. We propose a Bayesian semiparametric approach to estimate these two measures. The proposed estimators are based on the conditional distribution of the survival time given the biomarker and the empirical biomarker distribution. The conditional distribution is estimated by a linear dependent Dirichlet process mixture model. The resulting ROC curve is smooth as it is estimated by a mixture of parametric functions. The proposed c -index estimator is shown to be more efficient than the commonly used Harrell's c -index since it uses all pairs of data rather than only informative pairs. The proposed estimators are evaluated through simulations and illustrated using a lung cancer dataset.

Keywords

AUC; c -index; Concordance; Diagnostic test; Linear dependent Dirichlet process; Time-dependent ROC; ROC

1. Introduction

The receiver operating characteristic (ROC) curve is an important tool to examine the effectiveness of continuous diagnostic markers in predicting a medical outcome. In the context of a binary outcome (disease or no disease), the ROC curve displays the sensitivity (probability that the marker is larger than a cutoff for a diseased subject) versus 1-specificity (probability that the marker is larger than the cutoff for a non-diseased subject) for all possible cutoffs of the continuous marker. The diagnostic accuracy is often summarized by the area under the ROC curve (AUC), which can be interpreted as the probability that the test result from a randomly chosen diseased subject exceeds that for a randomly chosen nondiseased subject; $AUC = 1$ indicates perfect discrimination, while $AUC = 0.5$ indicates discrimination no better than chance. Another important measure to assess the discrimination potential of the marker is the c -index, which estimates the concordance between predicted and observed outcomes. When the outcome is binary, Hanley and McNeil (1982) have shown that the c -index is equivalent to the AUC.

In studies with censored survival outcomes, a marker is often measured at baseline while the disease (or event) may occur at any time thereafter. Thus, sensitivity, specificity and ROC curves are time-dependent, with $ROC(t)$ denoting the ROC curve at time t . Furthermore, survival outcomes are often censored, hence, most $ROC(t)$ estimators are built using standard survival methods. For example, Kaplan-Meier estimators are used by Heagerty et al. (2000), Chambless and Diao (2006), Uno et al. (2007) and Hung and Chiang (2010). Of these, Heagerty et al. (2000) used a weighted Kaplan-Meier estimator where the survival function given the marker is computed from the subset of subjects who have the nearest values of the marker (denoted as the NNE estimator). Hanley and McNeil (2003), Chambless and Diao (2006), and Song and Zhou (2008) considered the Cox model to develop an estimator, but this requires a proportional hazards assumption. Other semiparametric approaches include a marginal method (Cai et al., 2006) or a regression quantile method (Zheng and Heagerty, 2004). Bayesian nonparametric approaches have been proposed for binary outcomes (Erkanli et al., 2006; Branscum et al., 2008; Hanson et al., 2008; de Carvalho et al., 2011, 2013). In this paper, we will use a Bayesian semiparametric model, specifically, the linear dependent Dirichlet process (LDDP) mixture model, to estimate time-dependent ROC curves for censored survival outcomes. In the LDDP mixture model, the conditional distribution of the survival time given the marker is estimated by a mixture of linear models with a Dirichlet process (DP) prior on the random mixing measure. The proposed estimator for the ROC curve does not require a pre-fixed smoothing parameter as in the NNE estimator, and it produces a ROC curve that is automatically smoothed because it is a mixture of parametric functions.

In this paper, we also propose an estimator for the c -index. The c -index provides a global assessment of predictive accuracy of the survival model rather than focusing on the prediction at a fixed time. To deal with the censoring, Harrell et al. (1996) computed the c -index using usable patient pairs. Specifically, a patient pair is unusable if both patients died at the same time, or if one died and the other is still alive but has not been followed long enough to determine whether she will outlive the one who died. Harrell's c -index was studied extensively by Pencina and D'Agostino (2004). Unlike Harrell's c -index based on

informative pairs, Gönen and Heller (2005) and Begg et al. (2000) developed an estimator of concordance assuming the semiparametric Cox proportional hazards model that efficiently uses all patient pairs. In contrast, our c -index estimator incorporates all patient pairs using the LDDP model.

The rest of the article is organized as follows. In section 2 we introduce the LDDP mixture model and propose estimators for the time-dependent ROC curve and the c -index. In section 3 we present simulation studies. In section 4 we illustrate our method using a lung cancer dataset. Section 5 is the concluding discussion.

2. Proposed Bayesian Semiparametric Model

Let t_i be the survival time for subject i , and the minimum of t_i and c_i is observed where c_i is an independent censoring time, and $d_i = \mathbb{I}(t_i < c_i)$ is the censoring indicator. The design matrix for subject i is denoted by $\mathbf{z}_i = (1, \mathbf{z}_i)$, which includes an intercept and a baseline marker.

2.1 Linear Dependent Dirichlet Process Mixture Model

In a linear dependent Dirichlet process (LDDP) mixture model (De Iorio et al., 2009), the conditional density, $f(t_i|\mathbf{z}_i)$ is specified as

$$f(t_i|\mathbf{z}_i) = \int f(t_i|\mathbf{z}_i'\boldsymbol{\beta}, \sigma^2) dH(\boldsymbol{\beta}, \sigma^2) \\ H \sim DP(a, H^o) \quad (1)$$

The integrand $f(t_i|\mathbf{z}_i'\boldsymbol{\beta}, \sigma^2)$ is often called the kernel, and it has a location $\mathbf{z}_i'\boldsymbol{\beta}$ and scale σ . Since survival times are nonnegative, we choose a log normal density for the kernel, which is equivalent to a normal distribution for $\log(t_i)$. Hence, $f(\log(t_i)|\mathbf{z}_i)$ is a covariate-dependent mixture of normal densities with a Dirichlet process (DP) prior on random mixing measures. In the DP prior, $H^o = (H_\mu^o, H_{\sigma^2}^o)$ is a base measure, where H_μ^o is a Gaussian distribution and $H_{\sigma^2}^o$ is a probability measure on $\mathcal{R}^+$ (e.g., corresponding to an inverse-gamma distribution). Here, $a > 0$ is a concentration parameter characterizing prior precision. When a is large, H resembles H^o .

Under the stick-breaking representation of Sethuraman (1994), the conditional density in (1) can be written as

$$f(\log(t_i)|\mathbf{z}_i) = \sum_{k=1}^{\infty} \pi_k \phi(\log(t_i); \mathbf{z}_i'\boldsymbol{\beta}_k, \sigma_k^2) \quad (2)$$

where ϕ is a normal density with a mean $\mathbf{z}_i'\boldsymbol{\beta}_k$ and variance σ_k^2 . The mean of each component density changes linearly with $\mathbf{z}_i$. Each component k has a mixing weight of π_k , where $\pi_k = V_k \prod_{j < k} (1 - V_j)$, which is a probability formulated from a stick breaking process,

with $V_k \sim \text{Beta}(1, a)$, for $k = 1, \dots, \infty$. This representation characterizes the conditional density using an infinite mixture of linear models and the dependence on z_i 's is thereby determined by the model. Further details on DP and LDDP can be found in MacEachern (1999) and De Iorio et al. (2004). It is also important to note that the effect of the continuous biomarker defined in (2) implies that the expected log-transformed survival time changes linearly with the biomarker. If desired, the model can be extended by including additional terms (such as a quadratic term in z_i), or generalized by using a B-splines DDP mixture model as in de Carvalho et al. (2013).

In the LDDP mixture model, we assume that the conditional (on the biomarker) density can be well approximated by a mixture of normal densities. Furthermore, the use of normal densities leads to conditional conjugacy and greatly simplifies the computation. To complete model (2), we specify a conditionally conjugate base measure,

$$(\beta_k, \sigma_k^2) = N_2(\beta_k | \mu_\beta, \sum_\beta) \Gamma(\sigma_k^{-2} | a_0, b_0) \quad k=1, \dots, K, \quad (3)$$

with conjugate hyperpriors $\mu_\beta \sim N(\mu_0, \Sigma_0)$, and $\sum_\beta^{-1} \sim W(r, rR^{-1})$. Here, $\Gamma(\cdot | a, b)$ denotes a Gamma p.d.f with shape a and rate b ; $W(r, rR^{-1})$ denotes a Wishart distribution with degrees of freedom r and expectation R^{-1} .

We use the blocked Gibbs sampler of Ishwaran and James (2001) for posterior sampling. The blocked Gibbs sampler relies on truncation of the stick-breaking representation to a finite number of components, K . That is, $f(\log(t_i) | z_i) = \sum_{k=1}^K \pi_k f(\log(t_i); z_i' \beta_k, \sigma_k^2)$, where K is the upper bound on the number of components used for the approximation. The conditional density is thus estimated by a mixture of K linear models with mixing weights automatically determined by the data. The full conditional distributions needed for Gibbs sampling have simple conjugate forms. Once subjects are allocated to one of the components, a standard Gibbs sampling for the normal linear model proceeds within each component. Subjects with right-censored times are considered as missing data and are imputed from a right-truncated conditional distribution. The details of the Gibbs sampling algorithm are in Web Appendix A. The **DPpackage** in **R** (Jara et al., 2011) can also be used for the posterior estimation, which is based on the marginalization of the DP (MacEachern and Müller, 1998).

2.2 Estimation of time-dependent ROC curves

Heagerty and Zheng (2005) proposed several definitions of time-dependent ROC curves (denoted as $\text{ROC}(t)$) based on various definitions of sensitivity and specificity: 1) ROC defined with *Cumulative* sensitivity and *Dynamic* specificity is used to distinguish subjects having the event before a given time t and those having the event after the time t , 2) ROC defined with *Incident* sensitivity and *Dynamic* specificity is used to distinguish subjects having the event at a given time t and those having the event after the time t , and 3) ROC defined with *Incident* sensitivity and *Static* specificity is used to distinguish subjects

having the event at a given time t and those free of the event within a fixed follow-up period, $(0, t^*)$. In this section, we illustrate that these $ROC(t)$ can be estimated using parameters in the LDDP mixture model and the empirical distribution of the marker.

From now on, we assume that survival times have been log transformed. To estimate $ROC(t_l)$ at time point t_l , we first need to propose estimators for various definitions of sensitivity and specificity. In Heagerty et al. (2000), *Cumulative* sensitivity is defined as

$$Sens^C(t_l, c) = P(z > c | t \leq t_l) = \frac{P(t \leq t_l, z > c)}{P(t \leq t_l)},$$

where $c \in \mathbb{Z}$, and $\mathbb{Z}$ denotes the sample space of z . According to Akritas (1994), the numerator $P(t \leq t_l, z > c)$ is equal to

$$P(t \leq t_l, z > c) = \int_c^\infty F(t_l | z=s) dG_z(s), \quad (4)$$

where $F(t_l | z=s)$ is the conditional cumulative distribution function (cdf) given $z=s$ and G_z is the distribution of marker z . Based on the relationship in (4), we propose to estimate $Sens^C$ by

$$\widehat{Sens}^C(t_l, c) = \frac{\sum_{i=1}^n \hat{F}(t_l | z_i) I(z_i > c)}{\sum_{i=1}^n \hat{F}(t_l | z_i)}$$

The derivation of this estimator is given in Appendix A. Here, $\hat{F}(t_l | z_i)$ is calculated using parameters in the LDDP mixture model, which is given by

$$\hat{F}(t_l | z_i) = \sum_{k=1}^K \pi_k \Phi(t_l; z_i' \beta_k, \sigma_k^2),$$

Hence $F(t_l | z=s)$ is estimated by a mixture of normal cdfs with the k th component characterized by β_k , σ_k^2 , and a mixing probability π_k .

Incident sensitivity, $Sens^I(t_l, c) = P(z > c | t = t_l)$, is estimated by

$$\widehat{Sens}^I(t_l, c) = \frac{\sum_{i=1}^n \hat{f}(t_l | z_i) I(z_i > c)}{\sum_{i=1}^n \hat{f}(t_l | z_i)}$$

where the conditional density $\hat{f}(t_l | z_i)$ is estimated by a mixture of normal densities. That is,

$$\hat{f}(t_l | z_i) = \sum_{k=1}^K \pi_k \phi(t_l; z_i' \beta_k, \sigma_k^2).$$

Similarly, *Dynamic* specificity, $Spec^{\mathcal{D}}(t_l, c) = P(z_i | c | t > t_l)$, is estimated by

$$\widehat{Spec}^{\mathcal{D}}(t_l, c) = \frac{\sum_{i=1}^n \hat{S}(t_l | z_i) I(z_i \leq c)}{\sum_{i=1}^n \hat{S}(t_l | z_i)},$$

where $\hat{S}(t_l | z_i) = 1 - \hat{F}(t_l | z_i)$. That is, the specificity is estimated by a mixture of normal survival functions.

Static specificity, $Spec^{\mathcal{D}}(t^*, c) = P(z_i | c | t > t^*)$, is simply estimated by

$$\widehat{Spec}^{\mathcal{D}}(t^*, c) = \frac{\sum_{i=1}^n \hat{S}(t^* | z_i) I(z_i \leq c)}{\sum_{i=1}^n \hat{S}(t^* | z_i)},$$

where $(0, t^*)$ is a fixed follow-up period (static).

The $ROC(t_l)$ is produced by plotting the sensitivity vs. 1-specificity for all possible cutoffs of z . The area under the $ROC(t_l)$ curve, or $AUC(t_l)$, is calculated by integrating the $ROC(t_l)$ using the trapezoid rule. Based on the proposed sensitivity and specificity estimators, we can produce related time-dependent ROC curves and their corresponding $AUC(t_l)$. All these estimates are functions of the LDDP mixture model parameters and the empirical distribution of the marker. Using posterior draws from the MCMC sampling, we can easily derive the posterior mean of the estimator along with the credible interval.

2.3 Estimation of c-index

In many applications, a global concordance measure, regardless a time point t_l , is desirable. A c -index is a widely used measure of predictive discrimination. Estimation of the c -index is usually based on the proportion of patient pairs in which the predictions and outcomes are concordant (i.e., the predicted survival time is larger for the patient who lived longer). To deal with censored data, Harrell's c -index (Harrell et al., 1996) is constructed using only those pairs of patients in which at least one patient had died. Such pairs are called usable. A patient pair is unusable if both died at the same time, or one died and the other is still alive but has not been followed long enough to determine whether she will outlive the one who died. In this paper, we use all patient pairs (observed and censored) to estimate the c -index through the LDDP mixture model.

If higher values of z indicate higher risk of death, the c -index is defined as $c = P(t_i < t_j | z_i > z_j)$ for two randomly drawn observations (t_i, z_i) and (t_j, z_j) from the joint density $f(t, z)$. To estimate this requires the conditional distribution $f(t | z)$, which we obtain from the LDDP mixture model and also requires the distribution of z , for which we use the empirical distribution. Specifically, at each MCMC iteration, t_i is generated for subject i from a component-specific distribution, for example, $\log(\tilde{t}_i) \sim N(z_i / \beta_k, \sigma_k^2)$ if subject i is classified into the k th component. Then for each draw of t_i we have (t_i, z_i) , $i = 1, \dots, n$, without any

censored observation. From this we calculate c as the proportion of concordant pairs among all pairs in the sample, given by

$$\hat{c} = \frac{\sum_{(i,j); i < j} \mathbf{I}(z_i > z_j, \tilde{t}_i < \tilde{t}_j)}{\sum_{(i,j); i < j} \mathbf{I}(z_i > z_j)}. \quad (5)$$

Justification of this estimator is given in Appendix B.

Repeating this over MCMC iterations gives the posterior distribution of the c -index. We note that our approach of obtaining a Bayesian semiparametric estimates with an empirical distribution is similar to the work in Gustafson (2006) and Liu and Gustafson (2008), where population averaged inference is obtained by first fitting a model and then averaging over the covariates. The population distribution of z could also be obtained using the Bayesian bootstrap (Rubin, 1981), a DP mixture of normals (Lo, 1984; Escobar and West, 1995), or a Polya Tree model (Lavine, 1992). In this work, we used the empirical sample distribution of z to replace the unknown population distribution of z . Although this neglects the uncertainty about z , it is more analogous to Harrell's c -index.

3. Simulations

We carried out simulations to evaluate the performance of our proposed estimator. We focus on the $\text{ROC}(t)$ with *Cumulative* sensitivity and *Dynamic* Specificity (simulation results for $\text{ROC}(t)$ with *Incident* sensitivity and *Dynamic* Specificity can be found in the Web Appendix B).

Following the simulation setup in Pencina and D'Agostino (2004), we generated survival times from an exponential regression model $S(t|z) = S_0(t)^{\exp(\gamma z)}$, where $S_0(t)$ is the survival function of an exponential distribution with a rate of 0.025 and the baseline marker, $z \sim \mathcal{N}(0, 1)$. We set $\gamma = 2 \log(1.22)$ in **Scenario I** and $\gamma = 2 \log(2)$ in **Scenario II**, with sample size $n = 200$ or 400. By varying the last follow up time and censoring rate, the percentage of censoring is close to 20% or 40%.

Prior Specification—In the LDDP mixture model, we set stick-breaking weights $V_k \sim \text{Beta}(1, 1)$ for $k = 1, \dots, K$, where a was fixed to be one, which is a widely-used choice in applications. Ohlssen et al. (2007) suggests a value for K of $5 \times a + 2$, we used a slightly larger value of $K = 10$. Thus a maximum of 10 linear models were used to approximate the conditional density in (1). The sensitivity to the choice of a and K is investigated later in this section. For the normal-inverse gamma prior in (3), μ_0 is set to be a vector of zeros, $\Sigma_0 = 100I_2$, $r = 4$ and $R = I_2$, where I_2 denotes the 2×2 identity matrix. The normal prior for β_k is relatively vague since variances in Σ_0 are large and the degrees of freedom in the Wishart prior are very small. To specify a prior for σ_k^2 , we first obtained crude estimates of the mean and variance of σ_k^{-2} by fitting a log-normal model to the simulated data. Following strategies for setting hyperparameters (Dunson, 2010; De Iorio et al., 2009), we determined that $a_0 = 3$

and $b_0 = 1.5$ are a reasonable choices. We generated 10,000 Gibbs samples with a burn-in of 4,000 iterations, and we then used 6,000 iterations for posterior inference.

3.1 Evaluation of the c -index estimator

The true values of the c -index were calculated using numerical integration by Mathematica Research (2003), based on the formula for the exponential survival model. The true c -index is low in Scenario I (≈ 0.609) and high in Scenario II (≈ 0.778). In the simulation, we compared the proposed LDDP c -index to Harrell's c -index.

As presented in Table 1, absolute biases of the two estimators are similar. But the LDDP c -index is more efficient, as indicated by significantly reduced mean square errors in both scenarios with different sample sizes and different censoring rates. Although both estimators are nonparametric, the LDDP c -index used all patient pairs rather than only usable pairs as in Harrell's c -index, thereby making more efficient use of the censored data. This is also shown by a further reduction in RMSE when the censoring rate is increased from 20% to 40%. The coverage rates of the 95% uncertainty intervals are adequate for the LDDP methods, whereas for Harrell's method with 40% censoring, they can be quite a bit less than the nominal value.

3.2 Evaluation of ROC estimators

In each simulation, we estimated $AUC(t_j)$ and $ROC(t_j)$ at month $t_j = 5, 10, 20, 40, 50$, and 60. About 20% of the events occur before 5 months and 58 — 70% of the events occur before 60 months. The $ROC(t_j)$ is defined as the estimate of sensitivity at a fixed specificity of 0.25, 0.5, or 0.75. For each configuration, we used the Monte Carlo method to calculate true values of $ROC(t_j)$ and $AUC(t_j)$, by using one million data points of (t, z) simulated from the exponential survival model.

The performance of the $AUC(t_j)$ estimators is shown in Figure 1. As indicated by this figure, the bias of the LDDP estimator varies across t_j 's, with the smallest bias observed for middle time points. When marker z has a higher discrimination ability, such as in scenario II, the bias of the LDDP estimator is smaller than the Heagerty's estimator. Overall, the LDDP estimator is more efficient compared to the Heagerty estimator, as indicated by dramatically reduced mean square errors for all studied scenarios.

Figure 2 depicts the performance of the $ROC(t_j)$ estimators when $n = 400$. Similar to the performance of $AUC(t_j)$ estimator, the bias of the LDDP estimator is smaller when the marker has a higher discrimination ability, and the MSE is significantly lower for the LDDP estimator compared to the Heagerty's estimator. It is interesting to note that the bias of $ROC(t)$ is smaller for a smaller specificity. Additionally, the 95% coverage probabilities are in the range of 88% to 97%, with the majority of them between 90-95%.

We also did sensitivity analysis for the hyperparameters in the LDDP method. We considered a random a in the prior for stick-breaking weights and assumed $a \sim \Gamma(10, 1)$ and increased K to 51. The K of 51 was used based on Ohlssen et al. (2007)'s recommendation of $K = 5 \times a + 1$, with the expected value of a being 10. The data seemed to contain little information for estimating the parameter a , leading to similar performance of the $AUC(t_j)$

and the c -index compared to a fixed a of one. Also, a large K of 51 greatly increased computation effort without gain in the performance.

4. Application

In this section we illustrate the proposed method using data from 442 lung adenocarcinoma patients in a large retrospective, multi-site, blinded study. The median follow-up is 4 years and maximum follow-up is 17 years. During the follow-up, there are 54% death (about 23.3% and 42.3% of deaths occurred before 2 and 5 years respectively). In the published work of Li et al. (2014), the SAG mRNA expression level was shown to be a significant prognostic factor for patient overall survival. This result was obtained using a Cox proportional hazards model. In this application, we evaluated the discrimination ability of SAG on survival using time-dependent ROC curves and the c -index derived using the LDDP method.

Since the lung data is similar to the simulated data in section 3, all priors were the same as in the simulation study. We generated 15,000 Gibbs samples with a burn-in of 5,000 iterations, and we then used 10,000 iterations for posterior inference. The MCMC chains mixed well and the convergence of the MCMC sampling algorithm was further checked using several diagnostic procedures as recommended by Cowles and Carlin (1996).

The LDDP c -index is 0.55 (95% CI: 0.53,0.59). This c -index is small but it is significantly different from the null value of 0.5, and it is comparable to 0.57 (95% CI: 0.53,0.61) obtained by Harrell's method.

Figure 3 presents ROC curves at 2 and 5 years. The ROC curves using the LDDP method and Heagerty's method overlap, but the LDDP method yields a smooth ROC curve. The corresponding AUC estimates are very close from both methods (see Table 2), but the posterior standard deviations of the LDDP estimates are slightly smaller than the standard errors from Heagerty's method.

The LDDP method also allows us to derive conditional ROC curves, in the presence of covariates in addition to the biomarker. In this application, we found that the impact of SAG levels on survival differs between genders. Using the Cox proportional hazards model, the coefficient of SAG is 0.09 ($p = 0.31$) for males and 0.29 ($p = 0.005$) for females. We applied the LDDP method, with $\mathbf{z}_i = (1, \text{gender}, \text{gender} \times \mathbf{z}_i)$, where *gender* is coded as -1 and 1 for male and female, respectively and $\text{gender} \times \mathbf{z}_i$ is the interaction term. Then the Gibbs sampling proceeds as before. We observed that AUC^c for males and females are 0.48 (95% CI: 0.42,0.54) and 0.60 (95% CI: 0.53,0.65) at 2 years, respectively. This indicates that the discrimination ability of SAG is only observed for females. We also plot the estimated AUC^c over time in Figure 4. It is evident that AUC^c for females are consistently higher than males across different time points.

5. Discussion

In this paper, we provide a unified framework to evaluate the potential discriminatory ability of a marker for censored survival outcomes. We model the conditional density of the

survival time given the marker using the LDDP mixture model and derive estimators for the c -index and the time-dependent ROC curve. Both estimators show advantages over the existing estimators. Furthermore, the standard deviation and credible interval of the estimates do not rely on a large sample approximation and can be easily obtained from the posterior draws of the MCMC sampling.

The LDDP mixture model also enables us to derive a conditional c -index and ROC curve, which is useful when the discriminatory ability of the marker is different for different levels of a covariate. But the proposed approach can not be directly generalized to multiple markers. One way to deal with multiple markers is to estimate a risk score for each patient using a combination of markers under some working model, for example, under the proportional hazards model, and then consider the risk score as the new marker.

In this work, we used the blocked Gibbs sampler (Ishwaran and James, 2001) for posterior inference. Simulation results show that a mixture of at most of 10 linear models is sufficient for censored survival data with a sample size from 200 to 400.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully thank the referees for very useful comments that improved the presentation of the paper.

Appendix A

$$\text{Sens}^c(t_l, c) = \frac{P(t \leq t_l, z > c)}{P(t \leq t_l)} = \frac{\int_{-\infty}^{\infty} F(t_l | z=s) I(s > c) dG_z(s)}{\int_{-\infty}^{\infty} F(t_l | z=s) dG_z(s)} = \frac{\int_{-\infty}^{\infty} F(t_l | z=s) I(s > c) g_z(s) ds}{\int_{-\infty}^{\infty} F(t_l | z=s) g_z(s) ds} = \frac{E_z[F(t_l | z=s) I(s > c)]}{E_z[F(t_l | z=s)]}$$

If we use the empirical distribution of the marker, z , $\text{Sens}^c(t_l, c)$ can be estimated by

$$\widehat{\text{Sens}}^c(t_l, c) = \frac{n^{-1} \sum_{i=1}^n \hat{F}(t_l | z_i) I(z_i > c)}{n^{-1} \sum_{i=1}^n \hat{F}(t_l | z_i)} = \frac{\sum_{i=1}^n \hat{F}(t_l | z_i) I(z_i > c)}{\sum_{i=1}^n \hat{F}(t_l | z_i)}$$

Appendix B

$$c = P(t_i < t_j | z_i > z_j) = \frac{P(z_i > z_j, t_i < t_j)}{P(z_i > z_j)} = \frac{1}{P(z_i > z_j)} \iint \iint_{z_i > z_j, t_i < t_j} f(t_i, z_i) f(t_j, z_j) dt_i dt_j dz_i dz_j$$

Substituting the empirical distribution of z , we have

$$\hat{c} = \frac{1}{\sum_{(i,j;i < j)} I(z_i > z_j)} \sum_{(i,j;i < j)} \iint_{t_i < t_j} I(z_i > z_j) f(t_i | z_i) f(t_j | z_j) dt_i dt_j$$

Substituting random draws of t_i^* , $i = 1, \dots, n$, from the LDDP mixture model, the c -index can be estimated by

$$\hat{c} = \frac{\sum_{(i,j;i < j)} I(z_i > z_j, t_i^* < t_j^*)}{\sum_{(i,j;i < j)} I(z_i > z_j)}$$

References

- Akritis MG. Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*. 1994; 22:1299–1327.
- Begg BC, Craemer LD, Venkatraman ES, Rosai J. Comparing tumor staging and grading systems: a case study and a review of the issues, using thymoma as a model. *Statistics in Medicine*. 2000; 19:1997–2014. [PubMed: 10900448]
- Branscum AJ, Johnson WO, Hanson TE, Gardner IA. Bayesian semiparametric ROC curve estimation and disease diagnosis. *Statistics in Medicine*. 2008; 27:2474–2496. [PubMed: 18300333]
- Cai T, Pepe M, Zheng Y, Lumley T, Jenny N. The sensitivity and specificity of markers for event times. *Biostatistics*. 2006; 7:182–197. [PubMed: 16079162]
- Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine*. 2006; 25:3474–3486. [PubMed: 16220486]
- Cowles MK, Carlin BP. Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*. 1996; 91:883–904.
- de Carvalho VI, Jara A, Hanson TE, de Carvalho M. Bayesian nonparametric ROC regression modeling. *Bayesian Analysis*. 2013; 8:623–646.
- de Carvalho VI, Turkman AA, Nakas CT, Alonzo TA. Nonparametric Bayesian estimation of the three-way receiver operating characteristic surface. *Biometrical Journal*. 2011; 53:1011–1024. [PubMed: 22069202]
- De Iorio M, Johnson WO, Müller P, Rosner GL. Bayesian nonparametric non-proportional hazards survival. *Biometrics*. 2009; 65:762–771. [PubMed: 19210742]
- De Iorio M, Müller P, Rosner GL, MacEachern SN. An ANOVA model for dependent random measures. *Biometrics*. 2004; 99:205–215.
- Dunson, D. *Nonparametric Bayes applications to biostatistics*. Cambridge: Cambridge University Press; 2010.
- Erkanli A, Sung M, Costello EJ, Angold A. Bayesian semiparametric ROC analysis. *Statistics in Medicine*. 2006; 25:3905–3928. [PubMed: 16416403]
- Escobar MD, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*. 1995; 90:577–588.
- Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*. 2005; 92:965–970.
- Gustafson P. On robustness and model flexibility in survival analysis: Transformed hazard models and average effects. *Biometrics*. 2006; 63:69–77. [PubMed: 17447931]
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982; 143:29–36. [PubMed: 7063747]
- Hanley JA, McNeil BJ. Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics*. 2003; 30:93–111.

- Hanson T, Branscum AJ, Gardner I. Multivariate mixtures of polya trees for modelling ROC data. *Statistical Modelling*. 2008; 8:81–96.
- Hanson T, Kottas A, Branscum AJ. Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian non-parametric approaches. *JRSSC*. 2008; 57:207–225.
- Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*. 1996; 15:361–387. [PubMed: 8668867]
- Heagerty P, Lumley T, Pepe M. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000; 56:337–344. [PubMed: 10877287]
- Heagerty P, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics*. 2005; 61:92–105. [PubMed: 15737082]
- Hung H, Chiang C. Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics*. 2010; 38:8–26.
- Ishwaran H, James L. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*. 2001; 101:179–194.
- Jara A, Hanson TE, Quintana FA, Muller P, Rosner GL. Dppackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*. 2011; 40:1–25. [PubMed: 21796263]
- Lavine M. Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*. 1992; 20:1222–1235.
- Li H, Jia MTL, Wei D, Zhao Y, Chen G, Xu J, Zhao L, Thomas D, Beer DG, Sun Y. Inactivation of SAG/RBX2 E3 ubiquitin ligase suppresses KrasG12D-driven tumorigenesis. *The Journal of Clinical Investigation*. 2014; 124:835–846. [PubMed: 24430184]
- Liu J, Gustafson P. On average predictive comparisons and interactions. *International statistical review*. 2008; 76:419–432.
- Lo AY. On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*. 1984; 12:351–357.
- MacEachern S, Müller P. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*. 1998; 7:223–338.
- MacEachern, SN. ASA Proceeding of the Section on Bayesian Statistical Science. American Statistical Association; Alexandria, VA: 1999. Dependent nonparametric process.
- Ohlssen D, Sharples L, Spiegelhalter D. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Statistics in Medicine*. 2007; 26:2088–112. [PubMed: 16906554]
- Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*. 2004; 23:2109–2123. [PubMed: 15211606]
- Research, W. Inc Mathematica Version 5.0. Champaign, IL, U.S.A: 2003.
- Rubin DB. The Bayesian bootstrap. *The Annals of Statistics*. 1981; 9:130–134.
- Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica*. 1994; 4:639–650.
- Song X, Zhou XH. A semiparametric approach for the covariate specific ROC curve with survival outcome. *Statistica Sinica*. 2008; 18:947–965.
- Uno H, Cai T, Tian L, Wei L. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*. 2007; 102:527–537.
- Zheng Y, Heagerty P. Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics*. 2004; 5:615–632. [PubMed: 15475423]

Figure 1.

Performance statistics of AUC^c from the LDDP (lines with solid circles) and the Heagerty (lines with solid triangles) method with a sample size $n = 200$ or $n = 400$ and a censoring rate (cen) 20% or 40%. Scenario I and II have true c -index of 0.609 and 0.778 respectively. The x-axis shows the time point at which the AUC is calculated and the y-axis shows the performance statistics (Bias/MSE) of AUC estimates based on 500 simulated datasets. The Heagerty $AUC(t)$ was estimated using the *survivalROC.C* function in **R**.

Figure 2.

Performance statistics of ROC from the LDDP method (lines with solid circles) and the Heagerty (lines with solid triangles) method, with a sample size of $n = 400$, a censoring rate (cen) 20% or 40% and a specificity (q) of 25%, 50% or 75%. Scenario I and II have true c -index of 0.609 and 0.778, respectively. The x-axis shows the time point at which the ROC curve is calculated and the y-axis shows the performance statistics (Bias/MSE) of ROC estimates based on 500 simulated datasets.

Figure 3.

ROC curves for SAG measurements and survival in the lung cancer data. Panels (a)-(b) for ROC at 2 years and Panels (c)-(d) for ROC at 5 years. Solid lines are estimates using the LDDP method, with a band constructed using the pointwise 2.5% and 97.5% quantiles across simulations (presented in a shaded region). Dashed lines are estimates using the Heagerty method.

Figure 4. Lung cancer data results. Panel (a) shows the posterior mean estimates of $AUC^{\hat{\ell}}$ over time, and panel (b) depicts $ROC^{\hat{\ell}}$ at 2 years, using the proposed LDDP method.

Parameter estimation and performance statistics of the c-index based on 500 simulated datasets. The sqrtMSE denotes the square root of the mean square error (MSE) and RMSE= MSE of the LDDP/ MSE of Harrell. CP is the 95% coverage probability. Harrell's c-index, including the Mean and 95% confidence intervals, were estimated using the *concordance.index* function in **R**.

Table 1

	n	%cen	Method	Mean	Bias	sqrtMSE	RMSE	CP
Scenario I	200	20	LDDP	0.606	0.0006	0.0227	0.88	97.4
		20	Harrell	0.608	0.0015	0.0241		93.0
		40	LDDP	0.604	0.0026	0.0246	0.76	96.2
		40	Harrell	0.610	0.0028	0.0282		92.8
	400	20	LDDP	0.607	0.0000	0.0153	0.86	97.8
		20	Harrell	0.609	0.0017	0.0166		95.2
		40	LDDP	0.605	0.0024	0.0167	0.73	99.2
		40	Harrell	0.609	0.0020	0.0196		96.2
Scenario II	200	20	LDDP	0.775	0.0029	0.0168	0.87	93.6
		20	Harrell	0.779	0.0011	0.0180		91.8
		40	LDDP	0.773	0.0052	0.0183	0.63	94.4
		40	Harrell	0.788	0.0099	0.0230		88.4
	400	20	LDDP	0.775	0.0030	0.0113	0.92	95.6
		20	Harrell	0.778	0.0006	0.0118		94.8
		40	LDDP	0.773	0.0051	0.0124	0.58	95.6
		40	Harrell	0.787	0.0089	0.0163		87.8

Estimates of AUC^c and $AUC^{\mathcal{J}}$ from the Lung Cancer dataset. The standard error (SE) in the LDDP method was calculated from the posterior distribution and the SE in the Heagerty's method was calculated from the bootstrap approach as suggested in their paper.

Table 2

t_f (yrs)	Method	$AUC^c(t_f)$			$AUC^{\mathcal{J}}(t_f)$		
		Mean	SE	95% CI	Mean	SE	95% CI
2	LDDP	0.60	0.03	(0.55, 0.65)	0.56	0.02	(0.52, 0.59)
	Heagerty	0.61	0.03	(0.55, 0.67)	0.56	0.03	(0.51, 0.63)
5	LDDP	0.58	0.02	(0.53, 0.62)	0.53	0.02	(0.48, 0.57)
	Heagerty	0.57	0.03	(0.51, 0.63)	0.54	0.03	(0.50, 0.60)