Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2015 *Paper* 114

A Simple Method to Estimate the Time-dependent ROC Curve Under Right Censoring

Liang Li* Bo Hu†

Tom Greene‡

*The University of Texas M.D. Anderson Cancer Center, LLi15@mdanderson.org

†Cleveland Clinic

‡University of Utah

A Simple Method to Estimate the Time-dependent ROC Curve Under Right Censoring

Liang Li, Bo Hu, and Tom Greene

Abstract

The time-dependent Receiver Operating Characteristic (ROC) curve is often used to study the diagnostic accuracy of a single continuous biomarker, measured at baseline, on the onset of a disease condition when the disease onset may occur at different times during the follow-up and hence may be right censored. Due to censoring, the true disease onset status prior to the pre-specified time horizon may be unknown on some patients, which causes difficulty in calculating the time-dependent sensitivity and specificity. We study a simple method that adjusts for censoring by weighting the censored data by the conditional probability of disease onset prior to the time horizon given the biomarker and the observed censoring time. Our numerical study shows that the proposed method produces unbiased and efficient estimators of time-dependent sensitivity and specificity as well as area under the ROC curve, and outperforms several other published methods currently implemented in R packages.

A Simple Method to Estimate the Time-dependent ROC Curve Under Right Censoring

Liang Li[1], Tom Greene[2], Bo Hu[3]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, U.S.A. E-mail: LLi15@mdanderson.org
[2]Department of Population Health Sciences, University of Utah, Salt Lake City, Utah, U.S.A.
[3]Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio, U.S.A.

Abstract

The time-dependent Receiver Operating Characteristic (ROC) curve is often used to study the diagnostic accuracy of a single continuous biomarker, measured at baseline, on the onset of a disease condition when the disease onset may occur at different times during the follow-up and hence may be right censored. Due to censoring, the true disease onset status prior to the pre-specified time horizon may be unknown on some patients, which causes difficulty in calculating the time-dependent sensitivity and specificity. We study a simple method that adjusts for censoring by weighting the censored data by the conditional probability of disease onset prior to the time horizon given the biomarker and the observed censoring time. Our numerical study shows that the proposed method produces unbiased and efficient estimators of time-dependent sensitivity and specificity as well as area under the ROC curve, and outperforms several other published methods currently implemented in R packages.

keywords: Area under the ROC curve; Biomarker; Diagnostic medicine; Prediction accuracy; Receiver Operating Characteristic; Survival prediction.

1 Introduction

The Receiver Operating Characteristic (ROC) curve is widely used in medicine to quantify the diagnostic accuracy of a continuous biomarker on a disease condition (Pepe 2003; Zhou, Obuchowski and McClish, 2011). It visualizes the probability of both true and false positive diagnosis corresponding to a series of diagnostic rules, or cut off values in the context of a continuous biomarker, and helps researcher select the cut off with the desired diagnostic accuracy. The area under the ROC curve (AUC) summarizes both the probabilities of true and false positive diagnosis over all the possible cut off values into a single number between 0 and 1, which can be used as an overall index of the diagnostic accuracy of the biomarker. In many medical studies patients may start at baseline with no disease and the disease condition may develop at different time points later in the follow-up. In such situations, the binary classification of patients into

a true positive group with the disease and a true negative group without the disease may not be appropriate without taking time into consideration because the disease status can only be ascertained relative to a particular time point. For example, if the disease condition of scientific interest is death, then we may define the true positive group to be any patients who developed the disease within 3 years after the baseline and the true negative group to be those who did not develop the disease during that time period. The time point of 3 years is pre-specified as a scientifically relevant horizon, and other time horizons may be used. The binary classification ignoring the time-dependence of the disease condition would not be appropriate in this context because every patient will die eventually.

In a seminal paper, Heagerty, Lumley and Pepe (2000) extended the traditional ROC curve analysis for binary data to time to event data and proposed the time-dependent ROC curve for studies in which the disease status change over time. Let T denote the time from baseline to the occurrence of the disease on a continuous scale, and X be a continuous biomarker measured at baseline. The patient's disease status at time horizon τ is defined as $D(\tau) = 1\{T \leqslant \tau\}$, which equals 0 (no disease) or 1 (with disease). Without loss of generality, we assume that a higher value of X is associated with higher risk of the disease, and the decision rule of the diagnostic test is that if $X > c$, the patient is predicted to have the disease within the time interval $(0, \tau]$; if $X \leqslant c$, the patient is predicted to be disease free throughout this time interval. Similar to the traditional definition of sensitivity and specificity for the binary classification system, Heagerty et al (2000) defined the sensitivity and specificity at time τ as

$$\begin{array}{ll} P(X > c|T \leqslant \tau), & \text{sensitivity} \\ P(X \leqslant c|T > \tau), & \text{specificity.} \end{array} \tag{1}$$

These definitions of sensitivity and specificity are similar to those for the traditional binary case, except that the disease status is defined with respect to the time horizon τ. Therefore, these are called the time-dependent sensitivity and specificity, and the plot of sensitivity (i.e., true positive; on the vertical axis) and one minus specificity (i.e., false positive; on the horizontal axis) is called a time-dependent ROC curve at time horizon τ. The time-dependent sensitivity, specificity, and ROC curve are expected to vary with τ.

In the traditional binary case, the disease status and biomarker are known in the data set, and the sensitivity and specificity, and hence the ROC curve, can be easily calculated with the empirical probabilities. In the context of time-dependent ROC analysis, a challenge is that the time of disease occurrence T is not always observed due to right censoring. Suppose the data set includes n independent and identically distributed subjects, and the observed data for subject i are denoted by $\{X_i, Y_i, \delta_i\}$, where X_i is a continuous biomarker measured at baseline, Y_i is the observed time of disease onset or censoring, whichever is earlier, and δ_i is the censoring indicator, which equals to 1 if the disease onset is observed and 0 if the subject is censored. Let T_i and C_i be the true time of disease onset and censoring, then $Y_i = \min(T_i, C_i)$ and $\delta_i = 1\{T_i \leqslant C_i\}$.

We assume that C_i is independent of T_i conditionally on X_i, unless stated otherwise. The goal of the time-dependent ROC analysis is to use the data to estimate the time-dependent sensitivity and specificity in (1), and hence the time-dependent ROC curve and AUC, at a pre-specified τ, properly accounting for right censoring.

Heagerty et al (2000) proposed two estimation methods for time-dependent ROC curve analysis. The first method is based on the Bayes theorem, the empirical distribution of the biomarker, and the Kaplan-Meier estimator of the survival for the whole data set and for the subgroups with $X \leqslant c$ and $X > c$ respectively. This method is simple to implement but it does not guarantee that the sensitivity (or specificity) is monotone in c, and it does not apply to situations where the censoring time C depends on the biomarker. The second method, called the nearest neighbor estimation method, avoids these two drawbacks by estimating the bivariate distribution of the biomarker X and the true survival time T nonparametrically using kernel methods. A bandwidth is needed as a tuning parameter for the kernel. The sensitivity and specificity are calculated directly from the estimated bivariate distribution. Both methods are now available in the `survivalROC` package of `R` programming language (R Core Team, 2013). Several other methods have subsequently been proposed. Chambless and Diao (2006) proposed two methods. The first one uses recursive calculation over the ordered times of the events similar to the Kaplan-Meier approach to survival function estimation, and it does not guarantee the monotonicity or boundedness (between 0 and 1) of the specificity. The second method uses Bayes theorem to express the sensitivity and specificity in terms of the distribution function of the biomarker and the conditional distribution of the survival given the biomarker, and then estimates the latter using a Cox model. Song and Zhou (2008) studied a similar method but incorporated additional covariates in the Cox model, producing a covariate-specific time-dependent ROC analysis. Hung and Chiang (2010a) proposed another estimator based on inverse probability of censoring weighting (IPCW). Blanche, Dartigues and Jacqmin-Gadda (2013) proposed a conditional IPCW method that allows the censoring time to be dependent on the biomarker. Uno et al (2007), and Hung and Chiang (2010b) proposed similar IPCW based estimators when there are multiple biomarkers and the conditional relationship between survival and biomarkers are specified through a parametric or semi-parametric model. IPCW based methods require the estimation of the censoring distribution and weighting the uncensored data by the inverse of the censoring distribution at appropriate time points. An IPCW method is implemented in the `R` package `timeROC`. Blanche, Dartigues and Jacqmin-Gadda (2013) and Blanche, Latouche, and Viallon (2013) provided comprehensive reviews on the methods for time-dependent ROC analysis. The time-dependent ROC concept and methods have been extended to situations beyond random right censoring, including competing risks setting (Blanche et al 2013; Saha and Heagerty 2010), interval censoring (Li and Ma 2011), and semi-competing risks (Jacqmin-Gadda et al 2014). In this paper, we focus on the right censored data only. The goal of this paper is to describe a simple method for estimating the time-dependent sensitivity, specificity, and ROC curves (Sec-

tion 2), and apply simulations to compare this method with several existing methods that are available in `R` (Section 3). We further illustrate the method with a real data application in Section 4. Discussions are presented in Section 5.

2 Estimation

The sensitivity and specificity is easy to calculate in traditional ROC analysis because both the disease status D_i and biomarker X_i are available for every subject in the data set. In the time-dependent ROC analysis, the difficulty comes from the fact that the disease status at time horizon τ, $D_i(\tau)$, may be unknown for some subjects due to censoring. However, not all subjects have an undetermined disease status. There are four scenarios. First, if $Y_i > \tau$, we call that subject a control (or survivor, assuming without loss of generality that the disease condition represents death) at τ, and the disease status $D_i(\tau) = 0$. We assign a weight $W_i = 0$ to this subject. Second, if $Y_i \leqslant \tau$ and $\delta_i = 1$, we call that subject a case (or non-survivor) at time τ, and the disease status $D_i(\tau) = 1$. We assign a weight $W_i = 1$ to this subject. Third, if $Y_i = \tau$ and $\delta_i = 0$, then we know that $T_i > Y_i = \tau$ and the disease status $D_i(\tau) = 0$. The weight W_i of this subject is 0. Note that when the time is on a continuous scale, then theoretically the probability of this scenario is zero. Fourth, if $Y_i < \tau$ and $\delta_i = 0$, the disease status is unknown for this subject, but the probability that this subject is a non-survivor is $P(T_i \leqslant \tau | Y_i, X_i)$, and the probability that this subject is a survivor is $P(T_i > \tau | Y_i, X_i)$. For each subject with $Y_i < \tau$ and $\delta_i = 0$, we define the weight of this subject to be the probability of being a non-survivor:

$$W_i = P(T_i \leqslant \tau | Y_i, X_i) = 1 - \frac{S_T(\tau | X_i)}{S_T(Y_i | X_i)}$$

where $S_T(t|X) = P(T > t|X)$ denotes the conditional survival distribution of T given the biomarker X, which can be estimated using kernel weighted Kaplan-Meier method with a bandwidth h:

$$\hat{P}(T_i \geqslant t | X_i) = \prod_{s \in \Omega, s \leqslant t} \left\{ 1 - \frac{\sum_j K_h(X_j, X_i) 1(Y_j = s) \delta_j}{\sum_j K_h(X_j, X_i) 1(Y_j \geqslant s)} \right\} \tag{2}$$

where Ω is the set of distinct Y_i's with $\delta_i = 1$. This quantity can be calculated conveniently via the `R` function `survfit()` in the `survival` package, with the `weights` argument. The uniform kernel is used throughout the paper. The sensitivity and specificity can then be estimated in non-iterative, closed expression as:

$$\begin{aligned} \hat{P}(X_i > c | T_i \leqslant \tau) &= \frac{\sum_{i=1}^n W_i 1\{X_i > c\}}{\sum_{i=1}^n W_i} \\ P(X_i \leqslant c | T_i > \tau) &= \frac{\sum_{i=1}^n (1 - W_i) 1\{X_i \leqslant c\}}{\sum_{i=1}^n (1 - W_i)} \end{aligned} \tag{3}$$

In the special situation where there is no censoring, the disease status is known for every subject and $D_i(\tau) = W_i$ and W_i equals to either 0 or 1. The equations in (3) automatically reduce to the formula used in the traditional ROC analysis where the disease status is known at time τ; in that situation, the sensitivity is estimated by the empirical proportion of non-survivors with $X_i > c$ and specificity is estimated by the empirical proportion of survivors with $X_i \leqslant c$. The probability weight W_i can be interpreted as the probability that a subject is a case, or heuristically but equivalently, the fraction of the subject that is a case. Therefore, in the presence of censoring, the group of non-survivors includes not only those that are known to have developed the disease within time τ, but also "fractions" of those whose disease status is uncertain due to censoring prior to τ. Likewise, the group of survivors includes not only those that are disease free beyond time τ, but also "fractions" of those whose disease status is uncertain due to censoring prior to τ.

The following derivation provides theoretical justification for the heuristic arguments above in the case of sensitivity. The justification for the specificity estimator is similar.

$$\begin{aligned} P(X > c | T \leqslant \tau) &= \frac{E\left(1\{X > c\} \times 1\{T \leqslant \tau\}\right)}{E\left(1\{T \leqslant \tau\}\right)} \\ &= \frac{E\left\{1\{X > c\} E\left(1\{T \leqslant \tau\} | Y, X\right)\right\}}{E\left\{E\left(1\{T \leqslant \tau\} | Y, X\right)\right\}} \\ &= \lim_{n\to\infty} \frac{\sum_{i=1}^{n} 1\{X_i > c\} P(T_i \leqslant \tau | Y_i, X_i)}{\sum_{i=1}^{n} P(T_i \leqslant \tau | Y_i, X_i)} \end{aligned} \tag{4}$$

The time-dependent ROC curve can be calculated by plotting the sensitivity and 1-specificity for a range of c's and the AUC can be calculated by trapezoidal integration. The variance and confidence interval of sensitivity, specificity, and AUC can be estimated by bootstrap.

The estimators in (3) are not only simple to calculate with standard software, they also possess attractive properties. First, these estimators are both monotone in c, which is a desirable property, as elucidated in Heagerty et al (2000). In contrast, the Kaplan-Meier method in Heagerty et al (2000) and the recursive method in Chambless and Diao (2006) do not produce monotone estimators. Second, owing to the simple, quick and stable calculation, the bootstrap can be completed quickly. In contrast, while the bootstrap is recommended for the nearest neighbor method in Heagerty et al (2000), the computation is very lengthy due to the complexity of the algorithm for the point estimator. Third, the proposed method automatically accounts for the possible dependence between the censoring time C and the biomarker variable X. The Kaplan-Meier estimator of Heagerty et al (2000), the recursive estimator of Chambless and Diao (2006), and the inverse probability of censoring weighting estimators in Uno et al (2007) and Hung and Chiang (2010a) do not apply to this situation. While the inverse probability censoring weighting idea can be modified to adjust for such dependence (Blanche et al 2013), it requires modeling the conditional distribution of censoring time given the biomarker and this model is usually not

of direct scientific interest. Occasionally, the inverse probability weights may become excessively large due to a small estimated probability in the censoring distribution, causing instability in computation, particularly when the censoring rate is small or during automated bootstrap iteration. In contrast, the weights W_i in (3) are always confined between 0 and 1. Fourth, the nearest neighbor method of Heagerty et al (2000), perhaps the most widely used method in practice, involves the use of a bandwidth, and a practical guidance on how to choose the bandwidth is not yet available (Blanche et al 2013). In the numerical studies of Sections 3 and 4, we will show that the result of the nearest neighbor method is sensitive to the bandwidth choice. Notably, while the estimators in (3) also use a bandwidth, the results appear to be insensitive to the bandwidth. Further more, the proposed method can be easily modified and made to be invariant to monotone transformations of X, by using a span instead of a fixed bandwidth. A span is the proportion of subjects involved in the kernel estimation in (2).

The proposed method has connections to two existing methods. The model-based approach in Chambless and Diao (2006) is similar to (3). The difference is that they defined W_i as the conditional probability of $T_i \leqslant \tau$ given X_i for all subjects, but our W_i equals to the conditional probability of $T_i \leqslant \tau$ given X_i on those subjects with $Y_i < \tau$ and $\delta_i = 0$; W_i is either 0 or 1 for the rest of the subjects whose disease status is known from the data. In addition, the approach in Chambless and Diao (2006) is semi-parametric in the sense that the conditional distribution of T_i given X_i is modeled by a Cox model, while our method is nonparametric with the use of kernel. In the context of semi-competing risks data, Jacqmin-Gadda et al (2014) proposed an imputation estimator that is similar to the estimators in (3). However, their model between survival and biomarker is a parametric illness-death model, while the proposed method is nonparametric without explicit assumptions on the relationship between the biomarker and survival times. Additionally, in the context of the study in Jacqmin-Gadda et al (2014), they seemed to reach a conclusion that the IPCW method performs better than the imputation estimator. However, our general conclusion from the numerical study under the right censored data setting (Sections 3 and 4) is that the proposed method has better performance than all the methods in comparison, including the IPCW method. Since the right censored setting is the most widely encountered situation in practice, this finding is important for practical uses of the time-dependent ROC curve analysis. Furthermore, while the proposed method and other nonparametric method such as the nearest neighbor method (Heagerty et al 2000) both involve kernel estimation and a tuning parameter such as a bandwidth, we found that the proposed method is substantially less sensitive to the tuning parameter than the nearest neighbor method.

3 Simulation

We conducted detailed simulations to study the performance of the proposed method and compare it with three other methods: the Kaplan-Meier type

method (KM) and the nearest neighbor estimation method (NNE) proposed in Heagerty et al (2000) and implemented in the `R` package `survivalROC`, and the IPCW method studied by Hung and Chiang (2010a) and Blanche et al (2013), and implemented in `R` package `timeROC`. We simulate independent and identically distributed data assuming that X_i, $\log T_i$, and $\log C_i$ follow a trivariate normal distribution:

$$\begin{bmatrix} X_i \\ \log T_i \\ \log C_i \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ \mu_C \end{bmatrix}, \begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1\rho_2 \\ \rho_2 & \rho_1\rho_2 & 1 \end{bmatrix} \right)$$

In this model, we use ρ_1 to introduce different strengths of association between the biomarker X and the true survival time T, with higher biomarker indicating higher risk of the event ($-1 < \rho_1 < 0$), and use ρ_2 ($|\rho_2| < 1$) to introduce dependence between censoring time C and the biomarker variable X. The correlation between $\log T_i$ and $\log C_i$ is specified as $\rho_1\rho_2$, which ensures that $\log T_i$ and $\log C_i$ are conditionally independent given X_i. We use μ_C ($\mu_C = -0.5$ or 1) to control the censoring proportions, with the smaller value leading to higher proportion of censored data. The sample size n is 200 or 500. $\tau = 0.8$. In each scenario, 200 simulations were performed, and results were aggregated to study the percent bias and mean squared error in estimating the AUC.

Tables 1 and 2 show the performance of the four methods under every combination of n, μ_C, and ρ_1 with independent right censoring ($\rho_2 = 0$). The parameter to be estimated is the area under the ROC curve (AUC). We used a span of 0.1 in both the proposed and NNE methods, i.e., 10% of the data were used in the kernel estimation. The result shows that the proposed method, KM and IPCW all had negligible bias ($< 1\%$), but the NNE method had a small bias (usually $< 6\%$). The proposed method had the smallest MSE in almost all scenarios. The KM method appeared to be competitive in terms of bias, but we observed that sometimes it produced sensitivity and specificity estimates that were not monotone in the cut off values; this drawback was also observed by Heagerty et al (2000). The bias with the NNE was perhaps partially due to the suboptimal choice of the bandwidth, because we found that the result of this method is sensitive to the bandwidth choice. However, no practical guidance on the bandwidth selection has been proposed in the literature.

To further study this issue, we conducted additional simulations in Table 3, where we compared the proposed method with the NNE as the bandwidth (quantified by span) varied. The proposed method was clearly not sensitive to the bandwidth choice and had little bias, unless an unrealistically large span, such as 40% of the data, was used. The insensitivity increases with smaller percentage of censored data, as expected. The NNE method was much more sensitive to the bandwidth. This result offers an explanation of the relatively large bias associated with NNE method in Tables 1 and 2. A heuristic explanation of the insensitivity of the proposed method is as follows. First, for subjects who are not censored or those who are censored after the time horizon τ, their disease status is known for the purpose of calculating the sensitivity and specificity. The bandwidth plays a role only for those who are censored prior to time

τ, and this is a smaller proportion than the overall proportion of subjects who are censored. Second, the probability weight W_i for those subjects with $Y_i < \tau$ and $\delta_i = 0$ is defined through the ratio of two probabilities $S_T(\tau|X_i)/S_T(Y_i|X_i)$. If an inappropriately larger or smaller bandwidth causes a bias in $S_T(t|X_i)$, that bias usually goes in the same direction for different t's, especially in the case when $t = Y_i$ and $t = \tau$ are close. As a result, their biases in the weight may cancel to some extent.

Table 4 shows that the proposed method produced unbiased results when the censoring time is conditionally independent of the true survival given the the biomarker variable but there is correlation between the censoring time and biomarker. To put the percent bias numbers in perspective, we also reported the results from the KM method, which is known to be biased in such situations (Heagerty et al 2000).

4 Example

We illustrate the proposed method with a data set from a randomized placebo-controlled trial of the drug D-penicillamine (DPCA) for treating primary biliary cirrhosis (PBC) conducted at Mayo Clinic between 1974 and 1984. The data set has been used in many statistical publications (Murtaugh et al 1994; Therneau and Grambsch 2000; Fleming and Harrington 1991; Heagerty and Zheng 2005), and is available in the `survivalROC` package of `R`. The data include 312 subjects, of whom 125 died during the trial and the rest were right censored. We study the prognostic accuracy of a mortality risk prediction score developed from a Cox model with five baseline covariates. The score ranges from 3.74 to 11.25. Figure 1 shows the time-dependent ROC curves estimated from the proposed method (green), the NNE method (blue), and the IPCW method (red) at three time horizons: one year, three years, and six years. We did not include the KM method in the comparison because it produced non-monotone ROC curves. The curves produced by the IPCW method and the proposed method are very close, and the latter appears to be slightly smoother, perhaps due to the use of kernel smoothing. The NNE curve resulted in smaller sensitivity for the same estimated specificity. Both NNE and the proposed method used the same span, which equals $0.25n^{-0.2}$ where $n = 312$ is the sample size. This span was used for the example in the `survivalROC` documentation. We further examined the sensitivity of the proposed method and the NNE method to the span selection in Figure 2. Consistent with the simulation results in Table 3, the proposed method is much less sensitive to the span than the NNE method. This feature is important because currently there is little guidance in the literature on how to choose the span or bandwidth.

5 Discussion

In this paper, we studied a simple weighting method to estimate the time-dependent ROC curve nonparametrically under right censoring. Due to censoring, the true disease status of some subjects in the data may be uncertain. Consistent estimation of the time-dependent sensitivity and specificity can be achieved by simply weighting these subjects by their conditional probabilities of being either in the case or control groups, given the data. The proposed method is nonparametric in the sense that no parametric distributional assumption is made regarding the marginal, conditional, or joint distributions of the biomarker variable X and the time to event variable T. The method involves nonparametric estimation of the conditional distribution of T given X through kernel weighting with a bandwidth, similar to Heagerty et al (2000). However, the proposed method is much less sensitive to the choice of the bandwidth than the NNE method in Heagerty et al (2000). This is a desired property, given that the existing methods either use a parametric model (Chambless and Diao 2006; Jacqmin-Gadda et al 2014) to estimate the conditional distribution of T given X, or use nonparametric methods without proper guidance on how the bandwidth or span should be chosen. We compared the proposed weighting method with several popular methods that are currently implemented in `R`, and the proposed method demonstrated similar or improved performance in terms of bias and mean squared error. The proposed ROC curve estimator is simple to program, fast in computation, insensitive to bandwidth specification, and applicable when the censoring time is dependent on the biomarker variable. In addition, it proposes sensitivity and specificity probabilities that are monotone in the cut-off values, and automatically reduces down to the traditional sensitivity and specificity estimators when there is no censoring. It can also be made invariant to monotone transformation of the biomarker. `R` functions that implement the proposed method are available upon request.

Heagerty and Zheng (2005) proposed three definitions of the time-dependent sensitivity and specificity (cumulative/dynamic, incident/static, incident/dynamic). In this paper, we have focused only on the cumulative/dynamic definition because this is the most widely used definition in practice. Future work is needed to study the weighted estimation of sensitivity and specificity under other definitions.

Acknowledgements

This research is partially supported by NIH grants 5P30CA016672 and 5R01DK090046.

References

[1] Blanche, P., Dartigues, J.-F., Jacqmin-Gadda, H. (2013). Review and comparison of ROC curve estimators for a time-dependent outcome with marker-

dependent censoring. *Biometrical Journal*, 55, 687-704.

[2] Blanche, P., Latouche, A., Viallon, V. (2013). Time-dependent AUC with right-censored data: a survey. *Risk Assessment and Evaluation of Predictions*, 239-251, Springer, http://arxiv.org/abs/1210.6805.

[3] Chambless, L.E. and Diao, G. (2006). Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine* 25, 3474-3486.

[4] Fleming, T. and Harrington, D. (1991) Counting Processes and Survival Analysis. Wiley, New York.

[5] Heagerty, P.J., Lumley, T. and Pepe, M.S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56, 337-344.

[6] Heagerty, P.J. and Zheng, Y.Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* 61: 92-105.

[7] Hung, H. and Chiang, C. (2010a) Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics*, 38, 826.

[8] Hung, H. and Chiang, C. (2010b) Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data. *Scandinavian Journal of Statistics*, 37, 664679.

[9] Jacqmin-Gadda, H., Blanche, P., Chary, E., Touraine, C., Dartigues, J.-F. (2014). Receiver operating characteristic curve estimation for time to event with semicompeting risks and interval censoring. Statistical Methods in Medical Research, 1-17.

[10] Li, J. and Ma, S. (2011). Time-dependent ROC analysis under diverse censoring patterns. *Statistics in Medicine*, 30, 1266-1277.

[11] Murtaugh, P.A., Dickson, E.R., van Dam, G.M., Malinchoc, M., Grambsch, P.M., Langworthy, A.L., Gips, C.H. (1994) Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*, 20, 126-134.

[12] Pepe, M.S. (2003). The statistical evaluation of medical tests for classification and prediction. New York, NY: Oxford.

[13] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`

[14] Saha, P. and Heagerty, P.J. (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics*, 66, 999-1011.

[15] Song, X. and Zhou, X.-H. (2008). A semiparametric approach for the covariate specific ROC curve with survival outcome. *Statistica Sinica* 18, 947-965.

[16] Therneau, T. and Grambsch, P. (2000) Modeling Survival Data: Extending the Cox Model. Springer-Verlag, New York.

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

[18] Zhou, X.-H., Obuchowski, N.A., McClish, D.K. (2011). Statistical methods in diagnostic medicine (2nd ed). Wiley, New York.

Table 1: Simulation results comparing methods in estimating the AUC under independent right censoring ($\rho_2 = 0$). The sample size is 200. % censor: percent of data that are censored; % bias: percent bias; MSE: mean squared error.

ρ_1	% censor	method	true AUC	% bias	MSE ($\times 10^{-3}$)
-0.3	64	proposed	0.637	-0.43	3.05
-0.3	64	KM	0.637	0.23	3.13
-0.3	64	NNE	0.637	-3.0	3.20
-0.3	64	IPCW	0.637	0.71	3.10
-0.6	64	proposed	0.780	-0.31	2.24
-0.6	64	KM	0.780	0.45	2.46
-0.6	64	NNE	0.780	-4.4	3.64
-0.6	64	IPCW	0.780	0.45	2.36
-0.9	64	proposed	0.940	-0.01	0.389
-0.9	64	KM	0.940	0.58	1.14
-0.9	64	NNE	0.940	-3.6	1.94
-0.9	64	IPCW	0.940	0.28	0.445
-0.3	24	proposed	0.637	0.80	1.95
-0.3	24	KM	0.637	0.79	1.94
-0.3	24	NNE	0.637	-2.1	1.93
-0.3	24	IPCW	0.637	0.88	1.98
-0.6	24	proposed	0.780	-0.44	1.31
-0.6	24	KM	0.780	-0.46	1.32
-0.6	24	NNE	0.780	-4.6	2.81
-0.6	24	IPCW	0.780	-0.43	1.30
-0.9	24	proposed	0.940	0.06	0.216
-0.9	24	KM	0.940	0.02	0.232
-0.9	24	NNE	0.940	-3.7	1.73
-0.9	24	IPCW	0.940	0.07	0.220

Table 2: Simulation results comparing methods in estimating the AUC under independent right censoring ($\rho_2 = 0$). The sample size is 500. % censor: percent of data that are censored; % bias: percent bias; MSE: mean squared error.

ρ_1	% censor	method	true AUC	% bias	MSE ($\times 10^{-3}$)
-0.3	64	proposed	0.637	-0.20	1.21
-0.3	64	KM	0.637	0.36	1.23
-0.3	64	NNE	0.637	-3.7	1.53
-0.3	64	IPCW	0.637	0.44	1.46
-0.6	64	proposed	0.780	-0.25	0.781
-0.6	64	KM	0.780	0.23	0.841
-0.6	64	NNE	0.780	-5.1	2.42
-0.6	64	IPCW	0.780	0.15	0.892
-0.9	64	proposed	0.940	-0.16	0.160
-0.9	64	KM	0.940	0.28	0.407
-0.9	64	NNE	0.940	-4.5	2.16
-0.9	64	IPCW	0.940	0.12	0.176
-0.3	24	proposed	0.637	-0.30	0.651
-0.3	24	KM	0.637	-0.24	0.654
-0.3	24	NNE	0.637	-4.0	1.20
-0.3	24	IPCW	0.637	-0.22	0.663
-0.6	24	proposed	0.780	-0.06	0.437
-0.6	24	KM	0.780	-0.01	0.448
-0.6	24	NNE	0.780	-5.4	2.35
-0.6	24	IPCW	0.780	0.01	0.465
-0.9	24	proposed	0.940	-0.13	0.0836
-0.9	24	KM	0.940	-0.10	0.103
-0.9	24	NNE	0.940	-4.9	2.37
-0.9	24	IPCW	0.940	-0.12	0.0847

Table 3: Simulation results comparing the proposed method versus the nearest neighbor method (NNE) on the estimation of AUC under different bandwidths. $\rho_1 = -0.6$. $\rho_2 = 0$. The true AUC is 0.7803. The bandwidth is quantified by the span, which is the percent of data used in the kernel estimation. % bias: percent bias (%); MSE: mean squared error ($\times 10^{-3}$).

span	method	% bias	MSE	span	method	% bias	MSE
$n = 200$, censoring rate = 64%				$n = 200$, censoring rate = 24%			
0.05	proposed	-0.17	1.86	0.05	proposed	-0.39	1.29
0.05	NNE	-1.6	2.24	0.05	NNE	-2.2	1.82
0.1	proposed	-0.31	2.24	0.1	proposed	0.34	0.905
0.1	NNE	-4.4	3.64	0.1	NNE	-3.9	1.93
0.2	proposed	-0.81	1.88	0.2	proposed	-0.38	1.32
0.2	NNE	-8.6	6.23	0.2	NNE	-8.9	6.06
0.4	proposed	-2.9	2.42	0.4	proposed	-0.42	1.18
0.4	NNE	-14	13.2	0.4	NNE	-14	13.4
$n = 500$, censoring rate = 64%				$n = 500$, censoring rate = 24%			
0.05	proposed	-0.25	0.802	0.05	proposed	0.18	0.398
0.05	NNE	-2.4	1.24	0.05	NNE	-1.9	0.700
0.1	proposed	-0.53	0.793	0.1	proposed	0.09	0.409
0.1	NNE	-5.5	2.67	0.1	NNE	-5	2.04
0.2	proposed	-1.6	0.989	0.2	proposed	0.14	0.511
0.2	NNE	-11	7.69	0.2	NNE	-10	6.80
0.4	proposed	-4	1.78	0.4	proposed	-0.46	0.408
0.4	NNE	-16	16.9	0.4	NNE	-16	16.2

Table 4: Simulation results comparing methods in estimating the AUC when the censoring time C is correlated with the biomarker variable X but conditionally independent of the true survival time T given X. The true AUC is 0.7803. The sample size is 200. $\rho_1 = -0.6$. % censor: percent of data that are censored; % bias: percent bias; MSE: mean squared error.

n	ρ_2	% censor	method	% bias	MSE ($\times 10^{-3}$)
200	-0.4	66	proposed	-1.4	2.62
	-0.4	66	KM	7.4	7.20
	-0.4	21	proposed	-0.36	1.31
	-0.4	21	KM	0.42	1.39
	0.4	62	proposed	0.09	2.59
	0.4	62	KM	-8.0	5.18
	0.4	26	proposed	0.75	1.42
	0.4	26	KM	-0.82	1.27
500	-0.4	66	proposed	-0.62	1.06
	-0.4	66	KM	7.4	4.82
	-0.4	21	proposed	-0.14	0.419
	-0.4	21	KM	0.70	0.473
	0.4	63	proposed	-0.33	0.905
	0.4	63	KM	-8.5	4.93
	0.4	26	proposed	0.39	0.540
	0.4	26	KM	-1.1	0.573

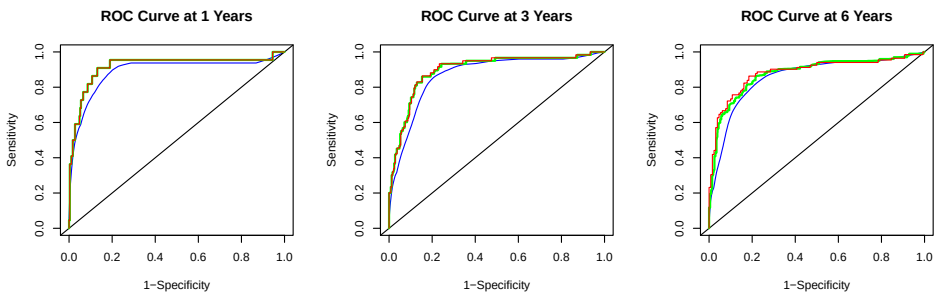

Figure 1: Time-dependent ROC curves estimated from the proposed method (green), the NNE method (blue), and the IPCW method (red) at three time horizons: one year, three years, and six years. The green and red curves are very close. We plotted the green curve with thicker lines to avoid overlap. The AUC of the three methods (proposed, NNE, IPCW) are $(0.918, 0.889, 0.918)$ for 1 year, $(0.898, 0.869, 0.898)$ for 3 years, and $(0.879, 0.856, 0.883)$ for 6 years

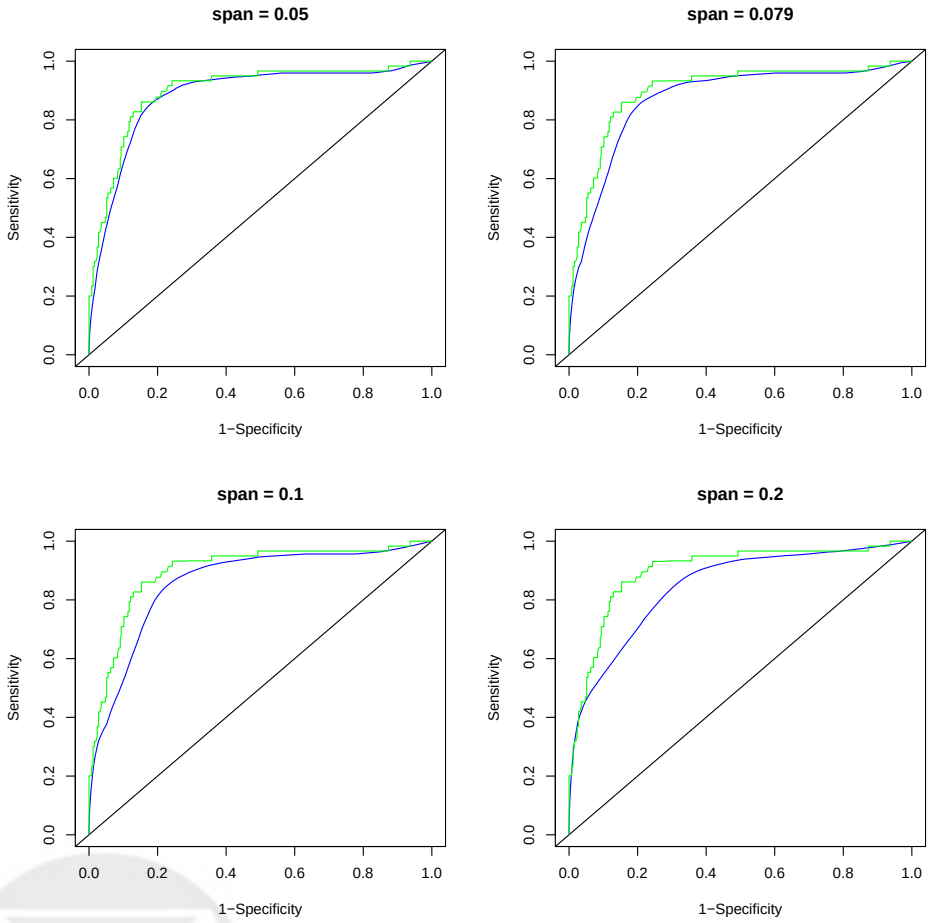

Figure 2: Time-dependent ROC curves estimated from the proposed method (green) and the NNE method (blue) using different span, which quantifies the percent of data used in each kernel calculation. The second span, 0.079, equals to $0.25n^{-0.2}$, where $n = 312$ is the sample size. The time horizon is 3 years. The proposed method is not sensitive to the choice of span.