

RESEARCH ARTICLE

A novel nonparametric time-dependent precision–recall curve estimator for right-censored survival data

Kassu Mehari Beyene¹  | Ding-Geng Chen^{1,2}  | Yehenew Getachew Kifle³

¹College of Health Solutions, Arizona State University, Phoenix, Arizona, USA

²Department of Statistics, University of Pretoria, Pretoria, South Africa

³Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland, USA

Correspondence

Ding-Geng Chen, Department of Statistics, University of Pretoria, Pretoria, South Africa.

Email: Ding-Geng.Chen@asu.edu

Funding information

South Africa National Research Foundation, Grant/Award Number: 114613

Abstract

In order to assess prognostic risk for individuals in precision health research, risk prediction models are increasingly used, in which statistical models are used to estimate the risk of future outcomes based on clinical and nonclinical characteristics. The predictive accuracy of a risk score must be assessed before it can be used in routine clinical decision making, where the receiver operator characteristic curves, precision–recall curves, and their corresponding area under the curves are commonly used metrics to evaluate the discriminatory ability of a continuous risk score. Among these the precision–recall curves have been shown to be more informative when dealing with unbalanced biomarker distribution between classes, which is common in rare event, even though except one, all existing methods are proposed for classic uncensored data. This paper is therefore to propose a novel nonparametric estimation approach for the time-dependent precision–recall curve and its associated area under the curve for right-censored data. A simulation is conducted to show the better finite sample property of the proposed estimator over the existing method and a real-world data from primary biliary cirrhosis trial is used to demonstrate the practical applicability of the proposed estimator.

KEYWORDS

precision–recall, prediction accuracy, right censored, risk score, survival

1 | INTRODUCTION

In precision health, prognosis often relates to the probability or risk of an individual developing a particular state of health or an outcome over a specific time, based on his or her clinical and nonclinical characteristics (Moons et al., 2009). Outcomes are often specific events, such as death, complications, and progression. Prognostic models are useful tools to estimate the risk that an individual in a particular health state will develop a particular health outcome. The estimated risk using prognostic model can give healthcare professionals an idea of the future course of patients' illness, so they can make decisions about treatment(s), such as deciding to start, stop, or change treatment(s). Statistical models, for example, survival models, are important tools to predict the probability that an individual will develop a particular state of health. As an example, the prognostic index of epithelial ovarian cancer (PIEPOC) is a risk score (or biomarker) derived using the Cox proportional hazards regression to predict the 5-year probability of overall survival for patients with advanced

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

epithelial ovarian cancer. This risk score classifies patients into low, intermediate, or high risk based on age, performance status, histological cell type, and tumor size (Mookerjee et al., 2007).

The predictive ability of a prognostic risk score should, however, be evaluated before it is used in clinical practice. In order to accomplish this, a risk score can be assessed for its ability to discriminate between low- and high-risk patients. The receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) are the two widely used tools to evaluate the discrimination ability of continuous biomarkers. The ROC curve is a plot of the true positive rate (the conditional probability that a diseased subject has a positive test) versus the false positive rate (the conditional probability that a healthy subject has a positive test) for all possible cutoff values. These measures are originally introduced for binary outcomes or known event status. In many practical situations, for example in time-to-event analysis, the disease outcomes, however, are time dependent. Therefore, using an ROC curve and AUC that varies with time is more appropriate. As a result, in the past few decades, many extensions have been introduced; see, for example, Heagerty et al. (2000), Heagerty and Zheng (2005), Etzioni et al. (1999), and Slate and Turnbull (2000). Following this, several time-dependent ROC curve and AUC estimation methods have been developed for various censoring mechanisms; see, for example, Heagerty et al. (2000), Heagerty and Zheng (2005), Li et al. (2018), Blanche et al. (2013a), Martínez-Cambor et al. (2016), Martínez-Cambor and Pardo-Fernández (2018), Beyene et al. (2019), Beyene and El Ghouh (2020, 2022), and the references given in these papers.

In recent years, precision–recall curves are becoming increasingly popular and has been shown to be a better alternative to the ROC curve for assessing the discriminatory ability of a continuous biomarker; see, for example, Ozenne et al. (2015), Saito and Rehmsmeier (2015), and Brodersen et al. (2010). A precision–recall curve is a plot of the true positive rate (also known as recall or sensitivity) against the positive predictive value also known as precision (the conditional probability that a subject with a positive test result actually has the disease) for all possible cutoff values. The precision–recall curve can be quantitatively summarized using the area under the precision–recall curve, also known as the average precision. The higher the under the precision–recall curve, the better the biomarker can classify subjects between classes. In spite of the fact that ROC curves and AUC are commonly used tools for assessing performance in a wide variety of applications, they overestimate the performances and may result in inaccurate conclusions when evaluating uncommon or rare disease biomarkers. In this case, the precision–recall curve and its summary measure are more meaningful as shown in Williams (2021), Sofaer et al. (2019), Saito and Rehmsmeier (2015), Ozenne et al. (2015), Brodersen et al. (2010), and Davis and Goadrich (2006). In spite of this advantage, Yuan et al. (2018) is the only research paper that has introduced the precision–recall curve and its associated area under the curve estimation for censored time-to-event data. This method, however, has two main limitations. First, the authors assumed that the event status is unknown for all the subjects under study without explicitly mentioning this. Second, this method assumes independent censoring, that is, the censoring time is independent of both the event time and the biomarker, which is unrealistic in many real-world biomedical studies. For the studies when this independent censoring assumption is not met, this method may produce an erroneous value for the area under precision–recall curve, which is out of the theoretical limit $[0, 1]$.

This paper is then aimed to address these limitations and propose a novel nonparametric estimation method for time-dependent precision–recall curve and the associated area under the curve, which would use all the available information efficiently. The rest of this paper is organized as follows. In the next section, we introduce some important notations and definitions, and propose estimation method for the true positive rate, the positive predictive value, the precision–recall curve, and the area under this newly proposed precision–recall curve. In Section 3, the finite sample performance of the proposed estimator is investigated through a simulation study followed by the illustration of this proposed novel method using a real data example in Section 4. Finally, we provide discussion and conclusion in Section 5.

2 | METHODS

In this section, we introduce some important notations and definitions, followed by the estimator for the time-dependent precision–recall curve and its summary measure.

2.1 | Notations and definitions

Let T be a nonnegative random variable denoting the time-to-event of interest (survival time or event time) and M denote a continuous biomarker that measured at baseline and its ability to predict an event T is to be evaluated. It is possible for

the biomarker to be either a single factor (for instance, prostate-specific antigen is used to predict prostate cancer risk) or a combination of risk factors (e.g., the score PIEPOC predicts death risk for patients with advanced epithelial ovarian cancer by combining age, performance status, histological cell type, and tumor size). Let us assume that the variable of interest T is subject to right-censoring, meaning that the exact survival time for some subjects is unknown except that it is greater than a certain value. This can happen for a variety of reasons, including when a subject leaves the study before a specific event occurs, or when the study ends before the event occurred. As a result the observed data set consists of $\{(Y_i, \Delta_i, M_i), i = 1, 2, \dots, n\}$, which are n independent copies of (Y, Δ, M) , where $Y = \min(T, C)$ is the observed time, $\Delta = I(T \leq C)$ is censoring indicator variable, and C is the censoring time, which is independent of T conditional on the biomarker M . For a given time of interest t , Heagerty et al. (2000) defined cases as subjects who experience the event before time t , that is, $T \leq t$, and controls as those who remain event-free through time t , that is, $T > t$. This is the most commonly used definition due to its clinical relevance (Blanche et al., 2013b; Lambert & Chevret, 2016). Assuming that the higher value of M is associated with higher risk of getting the event, for a given cutoff value m , the time-dependent true positive rate (TPR_t) is the conditional probability that M is greater than m given the event time T is less than or equal to t and the time-dependent positive predictive value (PPV_t) is the conditional probability that the event time is less than or equal to t given that the biomarker M is greater than m . Mathematically, these can be written as

$$\begin{aligned}\text{TPR}_t(m) &= P(M > m | T \leq t), \\ \text{PPV}_t(m) &= P(T \leq t | M > m),\end{aligned}\tag{1}$$

where $m \in (-\infty, \infty)$ is a fixed cutoff value. The corresponding time-dependent precision–recall curve is defined as a plot of the time-dependent true positive rate (or recall or sensitivity) versus the time-dependent positive predictive values (or precision) for all possible classification cutoff values, m . The time-dependent precision–recall (PR_t) curve can be quantitatively summarized by the area under the PR_t curve (AUPRC_t). As given in Yuan et al. (2018), this can be defined as

$$\text{AUPRC}_t = \int_{-\infty}^{\infty} \text{PPV}_t(s) d\text{TPR}_t(s).\tag{2}$$

As a measure of the discriminatory ability of a biomarker, AUPRC can be used to compare different biomarkers, where a higher value indicates a better biomarker performance. The maximum AUPRC value is 1, which corresponds to a perfect biomarker, means that the biomarker perfectly distinguishes between two classes.

2.2 | Review: Yuan method

In the case of censored time-to-event data, Yuan et al. (2018) proposed the only time-dependent precision–recall curve and area under the curve estimation method. Their approach is based on inverse probability censoring weighting (IPCW), and their estimators for TPR_t and PPV_t are given by

$$\begin{aligned}\widehat{\text{TPR}}_t(m) &= \frac{\sum_{i=1}^n \hat{w}_{ti} I(M_i > m) I(Y_i < t)}{\sum_{i=1}^n \hat{w}_{ti} I(Y_i < t)}, \\ \widehat{\text{PPV}}_t(m) &= \frac{\sum_{i=1}^n \hat{w}_{ti} I(M_i > m) I(Y_i < t)}{\sum_{i=1}^n I(M_i > m)},\end{aligned}\tag{3}$$

where $\hat{w}_{ti} = \frac{I(Y_j < t) \Delta_j}{\hat{G}(c)} + \frac{I(Y_j \geq t)}{\hat{G}(c)}$, and $\hat{G}(c)$ is a consistent estimator of the survival function of the censoring time, $G(c) = P(C \geq c)$. The time-dependent area under the precision–recall, which they call average precision (AP) can be written as

$$\widehat{\text{AP}}_t = \frac{1}{\sum_{j=1}^n \hat{w}_{tj} I(Y_j < t)} \sum_{j=1}^n \frac{I(Y_j < t) \hat{w}_{tj} \sum_{i=1}^n I(Y_i < t) \{I(M_i > M_j) + 0.5 \times I(M_i = M_j)\} \hat{w}_{ti}}{\sum_{i=1}^n \{I(M_i > M_j) + 0.5 \times I(M_i = M_j)\}}.\tag{4}$$

The main assumption of this approach is that the censoring time C is independent of both the event time T and the biomarker M . In case where this assumption is violated, which is the case in most of epidemiological studies, the Yuan method may produce an erroneous estimate. There is a strong consensus that the expected value of weights may not be 1 when the independence censoring assumption is violated, which was well documented in Howe et al. (2011), which studied the limitation of the IPCW approach. As a result, the precision–recall curve estimate may not necessarily be bounded in the square $[0, 1] \times [0, 1]$, and the time-dependent area under precision recall curve may not be contained within $[0,1]$. The Yuan method also assumed, without explicitly mentioning it, that all the subjects had unknown event status, which is unrealistic. Therefore, the objective of this study is to introduce a novel time-dependent precision–recall curve and the associated area under the curve estimation method that overcome the abovementioned limitations by using the available information efficiently.

2.3 | Proposed method

2.3.1 | Point estimation

Using the tower property of conditional expectations, or law of total probability, we first derive theoretical formulas for the time-dependent true positive rate and positive predictive value defined above, which is the basis for our proposal. Under the assumption that the event time T and then censoring time C are conditionally independent given the marker M , the time-dependent true positive rate given by (1) can be written as

$$\begin{aligned} \text{TPR}_t(m) &= \frac{P(M > m, T \leq t)}{P(T \leq t)} = \frac{E(I(M > m, T \leq t))}{E(I(T \leq t))}, \\ &= \frac{E\{E(I(M > m, T \leq t)|\Delta, Y, M)\}}{E\{E(I(T \leq t)|\Delta, Y, M)\}} = \frac{E\{I(M > m)W_t\}}{E\{W_t\}}, \end{aligned} \quad (5)$$

where W_t denotes a random variable $P(T \leq t|Y, \Delta, M)$, which can be written as

$$W_t \equiv P(T < t|M, \Delta, Y) = \left[1 - (1 - \Delta) \frac{S(t|M)}{S(Y|M)} \right] I(Y \leq t), \quad (6)$$

where $S(\cdot|M)$ denotes the conditional survival probability of T given the biomarker M . In this expression, the weight W_t is observed for all subjects except in the case when $Y < t$ and $\Delta = 0$. For the estimation of time-dependent ROC curve and the associated area under the curve, this weight was used and investigated in Beyene et al. (2019), Beyene and El Ghouch (2020), Li et al. (2018), and Martínez-Cambor et al. (2016). In the above weight function, the conditional survival function is unknown, which can be estimated from the observed data using the Beran estimator (Beran, 1981), see also Li et al. (2018), Beyene et al. (2019), and Beyene and El Ghouch (2020). The Beran estimator can be written as

$$\hat{S}(t|m) = \prod_{i=1}^n \left\{ 1 - \frac{I(Y_i < t, \Delta_i = 1)w_i(m)}{1 - \sum_{j=1}^n I(Y_j \geq Y_i)w_j(m)} \right\}, \quad (7)$$

where $w_i(m) = \frac{k((m-M_i)/h)}{\sum_{j=1}^n k((m-M_j)/h)}$ are Nadaraya–Watson weights (Nadaraya, 1964; Watson, 1964), $h \equiv h_n$ is a bandwidth, and k is a kernel function. In order to estimate bandwidth h , we propose to use the method presented by Sheather and Jones (1991). A similar suggestion is also made by Beyene and El Ghouch (2020).

Similarly, the theoretical formula for the time-dependent positive predictive value given in (1) can be defined as

$$\text{PPV}_t(m) = \frac{E\{E(I(M > m, T \leq t)|\Delta, Y, M)\}}{E(I(M > m))} = \frac{E\{I(M > m)W_t\}}{E(I(M > m))}. \quad (8)$$

From these, the estimator for the time-dependent true positive rate and positive predictive value can be defined by replacing the expectation in (5) and (8) with the empirical average. This means the empirical estimators for the TPR_t and PPV_t

are given by

$$\begin{aligned}\widehat{\text{TPR}}_t(m) &= \frac{\sum_{i=1}^n I(M_i > m) \hat{W}_{ti}}{\sum_{i=1}^n \hat{W}_{ti}}, \\ \widehat{\text{PPV}}_t(m) &= \frac{\sum_{i=1}^n I(M_i > m) \hat{W}_{ti}}{\sum_{i=1}^n I(M_i > m)},\end{aligned}\quad (9)$$

respectively, where $\hat{W}_{ti} = \left[1 - (1 - \Delta_i) \frac{\hat{S}(t|M_i)}{\hat{S}(Y_i|M_i)}\right] I(Y_i \leq t)$. Now, the time-dependent PR curve estimate can be obtained by plotting the estimated $\widehat{\text{PPV}}_t$ versus the estimated $\widehat{\text{TPR}}_t$ for all possible cutoff values.

The time-dependent area under precision–recall curve, as shown in Yuan et al. (2018), can be written as

$$\text{AUPRC}_t = \int_{-\infty}^{\infty} \text{PPV}_t(s) d\text{TPR}_t(s) = E_{M_1}(\text{PPV}_t(M_1)) = E(P(T_i < t | M_i > M_j, T_j < t)), \quad (10)$$

where M_1 denotes the biomarker for subjects with $T < t$. We can rewrite this as follows:

$$\begin{aligned}\text{AUPRC}_t &= E\left\{\frac{P(T_i < t, M_i > M_j, T_j < t)}{P(M_i > M_j, T_j < t)}\right\} \\ &= E\left\{\frac{E\{I(T_i < t, M_i > M_j, T_j < t)\}}{E\{I(M_i > M_j, T_j < t)\}}\right\} \\ &= E\left\{\frac{E(E\{I(T_i < t, M_i > M_j, T_j < t) | M_i, M_j, \Delta_i, \Delta_j, Y_i, Y_j\})}{E(E\{I(M_i > M_j, T_j < t) | M_i, M_j, \Delta_i, \Delta_j, Y_i, Y_j\})}\right\} \\ &= E\left\{\frac{W_j E(I(M_i > M_j) W_i)}{E(W_j) E(I(M_i > M_j))}\right\}.\end{aligned}\quad (11)$$

From this, the estimator for the time-dependent area under precision–recall curve can be given by

$$\widehat{\text{AUPRC}}_t = \frac{1}{\sum_{j=1}^n \hat{W}_{tj}} \sum_{j=1}^n \frac{\hat{W}_{tj} \sum_{i=1}^n \{I(M_i > M_j) + 0.5 \times I(M_i = M_j)\} \hat{W}_{ti}}{\sum_{i=1}^n \{I(M_i > M_j) + 0.5 \times I(M_i = M_j)\}}. \quad (12)$$

Note that the term $0.5 \times I(M_i = M_j)$ is included to handle tied marker values.

2.3.2 | Variance and confidence interval estimation

In order to estimate variance and/or approximate confidence intervals, we propose to use a nonparametric bootstrap method introduced by Efron (1979). This approach involves drawing B bootstrap samples of size n with replacement from the original data, and for each of this samples, we compute $\widehat{\text{AUPRC}}_b$, $b = 1, 2, \dots, B$. The empirical variance of the bootstrap estimates can be used to estimate the variance of $\widehat{\text{AUPRC}}_t$, which is given by

$$S_B^2 = B^{-1} \sum_{b=1}^B \left\{ \widehat{\text{AUPRC}}_b - B^{-1} \sum_{b=1}^B \widehat{\text{AUPRC}}_b \right\}^2. \quad (13)$$

The bootstrap estimates can also be used to approximate the $(1 - \alpha)100\%$ confidence interval for the theoretical AUPRC_t . To this end, the widely used is the percentile confidence interval that is constructed by $[\widehat{\text{AUPRC}}_B(\alpha/2), \widehat{\text{AUPRC}}_B(1 - \alpha/2)]$, where $\widehat{\text{AUPRC}}_B(\alpha)$ denotes the 100α th percentile of the B bootstrap estimates.

TABLE 1 Parameter values used to generate data and censoring proportion.

λ	γ	β	λ_C	Censored
1.5	2.0	-1.0	1.5	55%
1.5	2.0	-1.0	2.5	40%
1.5	2.0	-1.5	1.5	55%
1.5	2.0	-1.5	2.5	40%

The finite sample performance of the proposed time-dependent area under precision–recall curve and the variance and confidence interval estimation methods will be investigated in the next section.

3 | SIMULATION STUDY

In this section, we conduct a simulation study to investigate the performance of the proposed method under various scenarios with different censoring rate, sample sizes, and prediction times. Moreover, we also compare the proposed method with the competitor approach in the literature. In the upcoming subsections, we first present the data generation process before moving on to the simulation results discussions.

3.1 | Simulation setup

To generate the survival time, we considered a Weibull proportional hazards model given by

$$S(t|X) = \exp(-(t/\lambda)^\gamma \exp(X\beta)),$$

where $\lambda(\gamma)$ are the scale(shape) parameters, respectively, and β is the regression coefficient associated to the covariate X . From this, using the cumulative hazard inversion method, the survival time can be expressed as

$$T = \lambda^{-1}[-\log(U)\exp(-X\beta)]^{1/\gamma},$$

where U is a $[0,1]$ uniform distributed random variable and the covariate X is assumed to follow a standard normal distribution. The value of the scale and shape parameter is set to be 1.5 and 2, respectively.

The random censoring time C is generated from an exponential distribution with parameter λ_C , which is selected to achieve 40% and 55% censoring proportions. Moreover, we consider two regression coefficient values ($\beta = -1$ and $\beta = -1.5$), where the higher value corresponds to a stronger correlation between event time and biomarker. These parameters are summarized in Table 1.

For the estimation, we consider the prediction t corresponding to the quartile values of the survival time T (i.e., $t = Q_1$, $t = Q_2$, and $t = Q_3$) and the biomarker is defined as $M = \exp(-((T/1.5)^2)\exp(\beta'X))$. To compute the performance measures, we generate $N = 1000$ independent samples with two sizes ($n = 250$ and $n = 500$).

As a performance criterion, we consider the absolute bias (Bias), mean square error (MSE), and standard deviation (SD), which, respectively, are defined as

$$\begin{aligned} \text{Bias} &= |N^{-1} \sum_{s=1}^N \widehat{\text{AUPRC}}_{ts} - \text{AUPRC}_t|, \\ \text{MSE} &= N^{-1} \sum_{s=1}^N (\widehat{\text{AUPRC}}_{ts} - \text{AUPRC}_t)^2, \\ \text{SD} &= \sqrt{(N-1)^{-1} \sum_{s=1}^N (\widehat{\text{AUPRC}}_{ts} - \widehat{\text{AUPRC}}_t)^2}, \end{aligned}$$

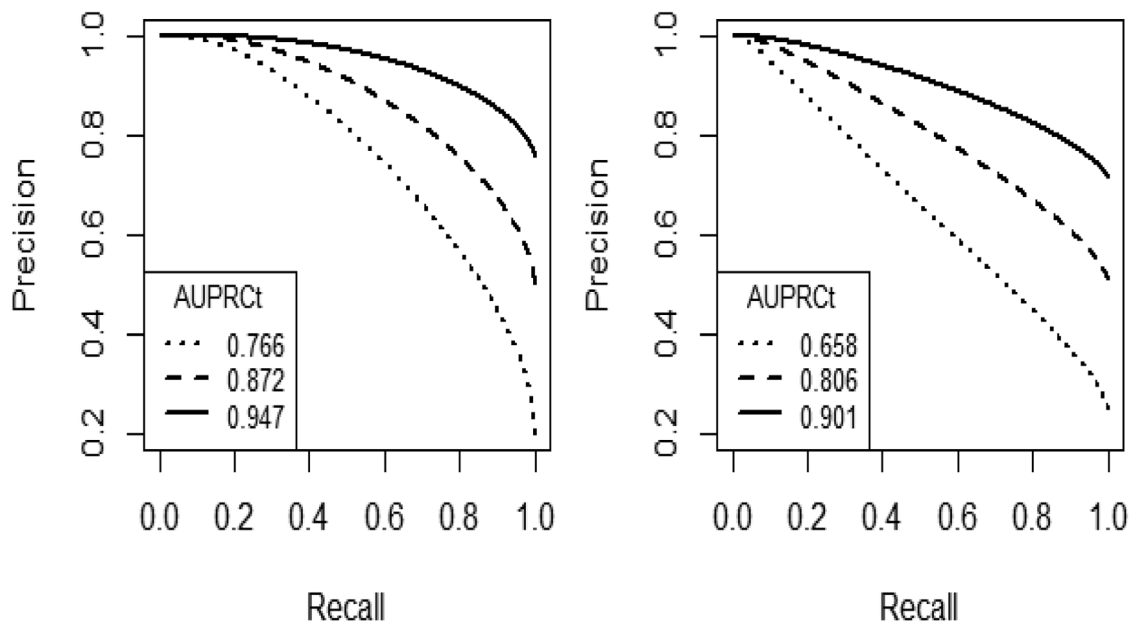


FIGURE 1 The true precision-recall curves and the corresponding area under the PR_t curve (AUPRC) values of the data generating process with $\beta = -1$ (left column) and $\beta = -1.5$ (right column) for the prediction time $t = 0.6$ (dotted line), $t = 1.2$ (dashed line), and $t = 2.0$ (solid line).

where $I(\cdot)$ is an indicator function, $\widehat{AUPRC}_{t,s}$ is the estimate obtained from the s^{th} simulated data. Here, the true $AUPRC_t$ is computed from simulated data $\{(T_i, M_i), i = 1, 2, \dots, 2 \times 10^6\}$, using the formula given in (12) with $W_{ti} = I(T_i < t)$.

The true time-dependent precision-recall curve and the associated area under the precision-recall curve are presented in Figure 1. As seen from the figure, both the precision-recall curve and the area precision-recall curve increases (decreases) with prediction time t (β).

3.2 | Simulation results

3.2.1 | Evaluating and comparing finite sample performance

In this section, we present and discuss the results obtained from simulations conducted to evaluate the performance of the proposed method on finite samples. Furthermore, we compare our method with the approach proposed by Yuan et al. (2018) and implemented in the R package `APtools` Cai et al. (2018). Table 2 presents the empirical SD, the absolute bias (Bias), and the MSE to evaluate the performance of the proposed method and compare it with the estimator proposed by Yuan et al. (2018), hereafter denoted by (Yuan). We computed the performance measures from 1000 simulated samples with sample sizes $n = 250, 500$, censoring proportions 40% and 55%, regression coefficients $\beta = -1.0, -1.5$ and prediction times $t = 0.6, 1.2, 2.0$. From the results, as expected, the SDs, biases and MESs of the proposed increases with increase in censoring rate. In contrast, the performance of the proposed estimator improves with increase in sample size, and prediction time. This is true for all simulation settings.

Compared to the Yuan method, the proposed method, in general, has smaller MSE and the magnitude of the MSE is less than 0.005 which is negligible. Similarly, the SD of the proposed method is smaller than the Yuan method with comparable bias as shown in Table 2. In Figure 2, we provide the boxplots of the time-dependent area under precision-recall curve estimates obtained from both the proposed approach and the Yuan method. This figure indicate that the performance of the proposed estimator improves with prediction time and sample size, and deteriorate with censoring rate. Furthermore, in general, the proposed method performed better than the Yuan approach, which is consistent with the result given in Table 2.

We would like to emphasis an interesting finding from this figure: the Yuan approach resulted in estimated $AUPRC_t$ values that exceed 1, which is outside the theoretical boundary $[0,1]$. These erroneous values are due to the fact that the

TABLE 2 Bias ($\times 100$) and mean square error (MSE) ($\times 1000$) of the proposed approach and the Yuan method computed using sample sizes (n), censoring rates (cens), and prediction time (t).

t	cens	n	TRUE	Proposed				Yuan			
				Estimate	SD	Bias	MSE	Estimate	SD	Bias	MSE
$\beta = -1.0$											
0.6	40	250	0.766	0.752	0.059	1.424	3.653	0.776	0.072	0.984	5.267
1.2	40	250	0.872	0.866	0.029	0.635	0.891	0.877	0.048	0.538	2.362
2.0	40	250	0.947	0.946	0.016	0.117	0.262	0.954	0.039	0.744	1.560
0.6	55	250	0.766	0.744	0.061	2.196	4.212	0.782	0.084	1.552	7.255
1.2	55	250	0.872	0.863	0.033	0.890	1.141	0.884	0.062	1.169	3.969
2.0	55	250	0.947	0.945	0.020	0.228	0.397	0.961	0.053	1.388	3.033
0.6	40	500	0.766	0.756	0.042	1.003	1.895	0.771	0.052	0.519	2.697
1.2	40	500	0.872	0.868	0.020	0.448	0.424	0.875	0.033	0.252	1.077
2.0	40	500	0.947	0.945	0.012	0.166	0.142	0.951	0.028	0.351	0.791
0.6	55	500	0.766	0.749	0.045	1.669	2.323	0.773	0.060	0.654	3.664
1.2	55	500	0.872	0.865	0.023	0.659	0.561	0.875	0.042	0.344	1.782
2.0	55	500	0.947	0.945	0.015	0.221	0.222	0.952	0.038	0.464	1.433
$\beta = -1.5$											
0.6	40	250	0.658	0.650	0.064	0.780	4.203	0.668	0.074	1.024	5.521
1.2	40	250	0.806	0.800	0.038	0.643	1.449	0.811	0.051	0.536	2.626
2.0	40	250	0.901	0.899	0.025	0.245	0.610	0.909	0.044	0.751	1.978
0.6	55	250	0.658	0.646	0.066	1.223	4.567	0.673	0.083	1.515	7.040
1.2	55	250	0.806	0.797	0.043	0.809	1.919	0.816	0.064	1.043	4.246
2.0	55	250	0.901	0.897	0.031	0.416	0.950	0.914	0.058	1.255	3.482
0.6	40	500	0.658	0.651	0.045	0.701	2.097	0.662	0.050	0.420	2.537
1.2	40	500	0.806	0.802	0.026	0.425	0.717	0.809	0.035	0.252	1.225
2.0	40	500	0.901	0.898	0.019	0.285	0.361	0.903	0.031	0.210	0.970
0.6	55	500	0.658	0.647	0.048	1.119	2.394	0.664	0.057	0.610	3.289
1.2	55	500	0.806	0.800	0.030	0.607	0.918	0.810	0.043	0.395	1.892
2.0	55	500	0.901	0.898	0.022	0.283	0.507	0.906	0.040	0.477	1.658

simulated data are generated under the realistic assumption that the event time T and the censoring C are independent given the marker M , which is different and more realistic than the Yuan approach where it was assumed that C is independent of both T and M . As indicated in Section 2.2, the IPCW-based estimator of the Yuan approach may not yield weight estimates that sum to one when the independence censoring assumption is not met (Howe et al., 2011). As a result, the precision-recall curve estimates may not necessarily be bounded in the square $[0, 1] \times [0, 1]$, and the $AUPRC_t$ estimates may not be contained within $[0, 1]$. From the simulation study, in general, we can conclude that the proposed method performs better than the Yuan method. In our simulations, we also examined the performance of the proposed method under conditions where the censoring time is marker-dependent. Across all the scenarios examined, our method consistently outperformed the Yuan approach, exhibiting smaller SD, bias, and MSE values. For brevity, we have omitted these results.

3.2.2 | Evaluating variance and confidence interval estimation method

In this section, we investigate the performance of the proposed bootstrap-based variance and confidence interval estimation method. Table 3 shows the empirical standard deviation (ESD), the average bootstrap standard deviation (ASD), the average width (AW) and coverage probability (CP) of the 95% bootstrap confidence intervals computed using the percentile approach for the proposed method. The ESD of the time-dependent area under precision-recall curve is the SD of the estimated $AUPRC_t$ obtained from the simulated data. The ASD of the estimator is computed as the average of the bootstrap standard deviations obtained from the simulated data. The AW is the average difference of the upper and lower

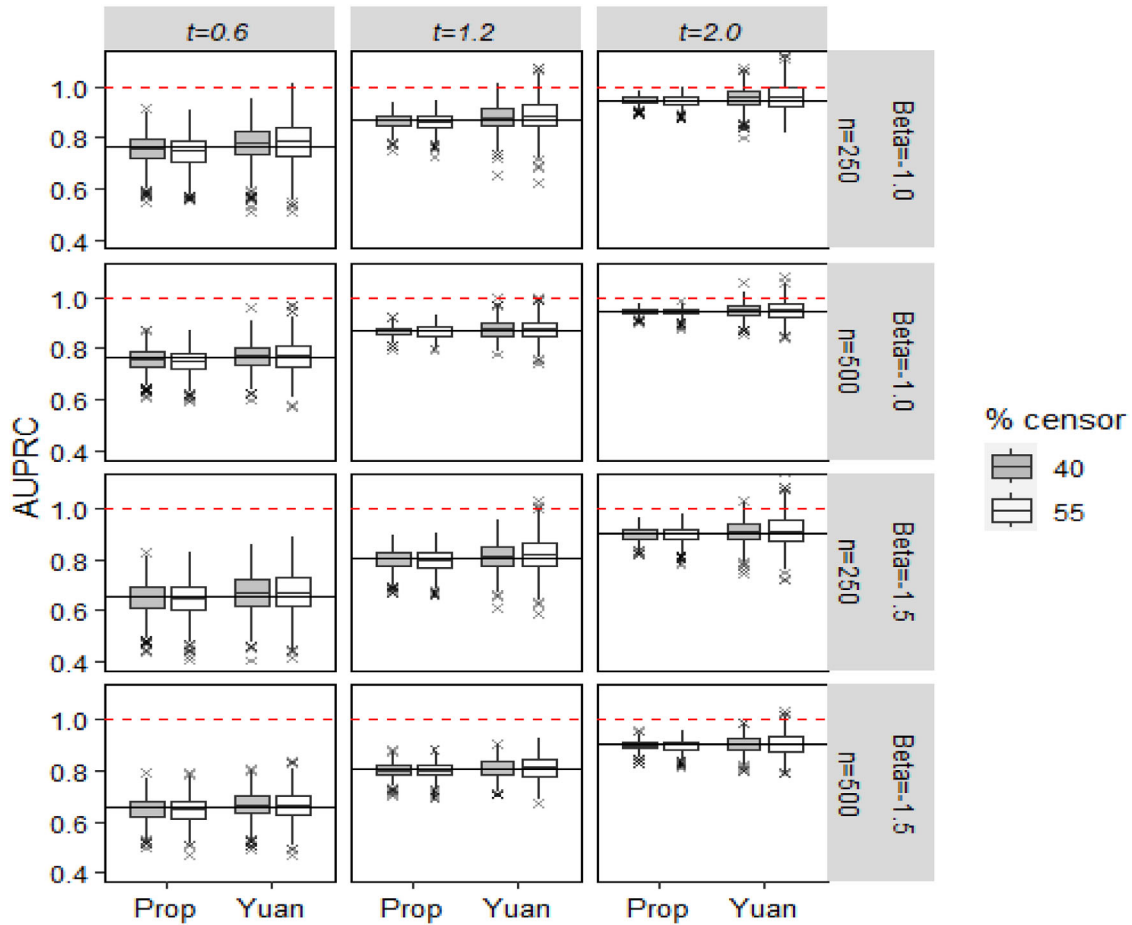


FIGURE 2 Boxplots for the estimated $AUPRC_t$ obtained using the proposed method (Prop) and the Yuan approach (Yuan). True values are indicated by horizontal solid black lines.

TABLE 3 The empirical standard deviation (ESD), the average bootstrap standard deviation (ASD), the average width (AW), and coverage probability (CP) of the 95% bootstrap confidence intervals for different sample sizes (n), censoring proportions (cens), t , and β values computed using the percentile approach with the proposed method.

t	cens	n	$\beta = -1$				$\beta = -1.5$			
			ESD	ASD	AW	CP(%)	ESD	ASD	AW	CP(%)
0.6	40	250	0.059	0.058	0.228	0.948	0.064	0.062	0.241	0.940
1.2	40	250	0.029	0.029	0.114	0.954	0.038	0.038	0.147	0.950
2.0	40	250	0.016	0.016	0.061	0.924	0.025	0.025	0.098	0.941
0.6	55	250	0.061	0.062	0.239	0.949	0.066	0.064	0.251	0.941
1.2	55	250	0.033	0.033	0.127	0.951	0.043	0.042	0.163	0.941
2.0	55	250	0.020	0.019	0.075	0.924	0.031	0.031	0.120	0.939
0.6	40	500	0.042	0.041	0.161	0.939	0.045	0.044	0.173	0.943
1.2	40	500	0.020	0.021	0.080	0.954	0.026	0.027	0.104	0.946
2.0	40	500	0.012	0.011	0.044	0.928	0.019	0.018	0.071	0.943
0.6	55	500	0.045	0.044	0.170	0.932	0.048	0.046	0.181	0.939
1.2	55	500	0.023	0.023	0.090	0.955	0.030	0.030	0.115	0.934
2.0	55	500	0.015	0.014	0.054	0.936	0.022	0.022	0.085	0.945

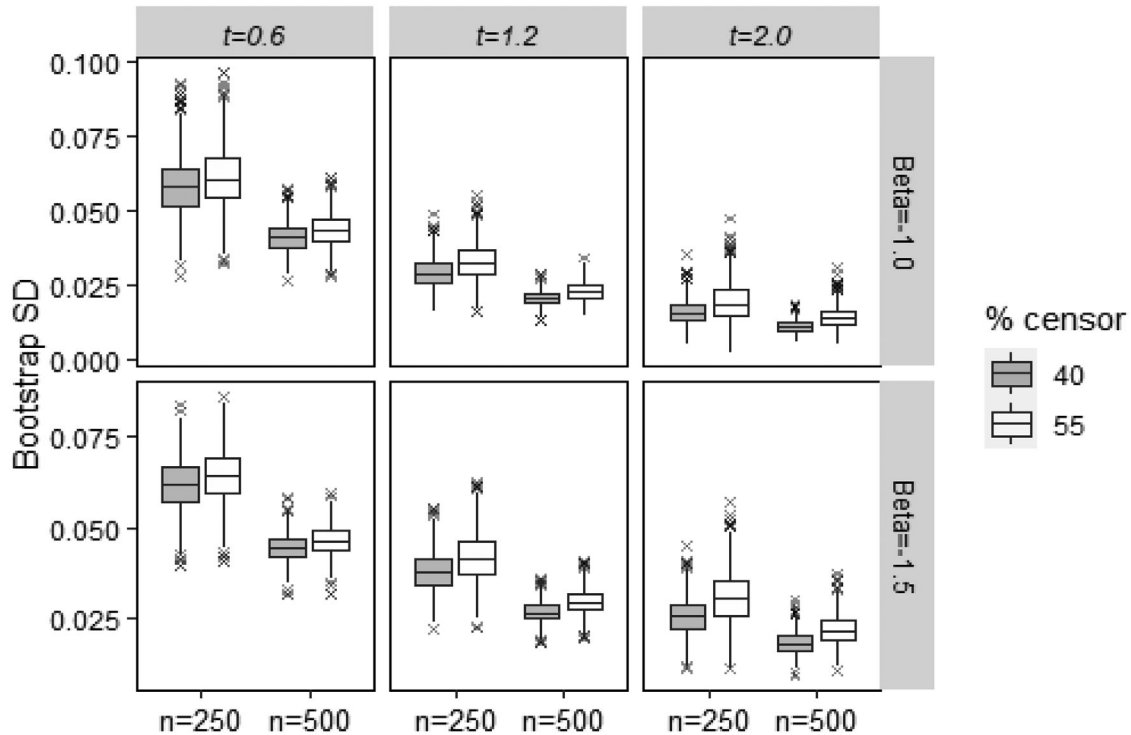


FIGURE 3 Boxplots for the estimated bootstrap standard deviations (SD) obtained using the proposed method.

limit of the bootstrap confidence interval. Finally, the CP is the proportion of bootstrap confidence interval that included the true value. The CP and AW are important measures to evaluate the validity and precision, respectively, of the proposed confidence interval estimation method.

From the simulation results presented in Table 3, both the ESD and ASD increase with increase in censoring rate and both decrease with increase in sample size and prediction time. Furthermore, the values of ESD and ASD, in general, are very close to each other indicating the consistency of the variability estimate obtained from the bootstrap method. Generally, the AWs of the proposed bootstrap confidence interval increase with sample size and decrease with censoring rate. In addition, for all simulation settings, the proposed percentile-based bootstrap confidence interval estimation method, in general, provides CPs that are very close to the nominal value of 0.95. Figures 3 and 4 show the boxplots of the bootstrap-based SD estimates and the width of 95% bootstrap confidence intervals of $AUPRC_t$, respectively. From the figures, the performance of both the proposed variance and confidence interval estimation methods improve (deteriorate) with sample size (censoring), which is consistent with the result in Table 3. Based on these, we can confidently conclude that the proposed bootstrap method makes good approximations of both variances and confidence intervals.

4 | REAL DATA APPLICATION

4.1 | Mayo PBC data

The data used to illustrate the proposed method is obtained from a randomized placebo-controlled trial of D-penicillamine for treating primary biliary cirrhosis (PBC) conducted between 1974 and 1984 at the Mayo Clinic. In this trial, there were 312 patients randomly assigned to receive D-penicillamine ($n = 158$) or placebo ($n = 154$), of whom 125 experienced the event of interest (death) during the follow-up period (Fleming & Harrington, 2011). The objective of this study was to develop models to predict the survival of patients with PBC disease. Heagerty and Zheng (2005) used a Cox proportional hazards model to derive two risk scores from these data. The first risk score (M_1) is calculated based on five covariates: log(bilirubin), albumin, log(prothrombin time), edema, and age. The second risk score (M_2) is derived from all the covariates, except log(bilirubin). These data are available in the R package `survivalROC` (Heagerty & packaging by Paramita Saha-Chaudhuri, 2022) and further details about these data can be found in Heagerty and Zheng (2005).

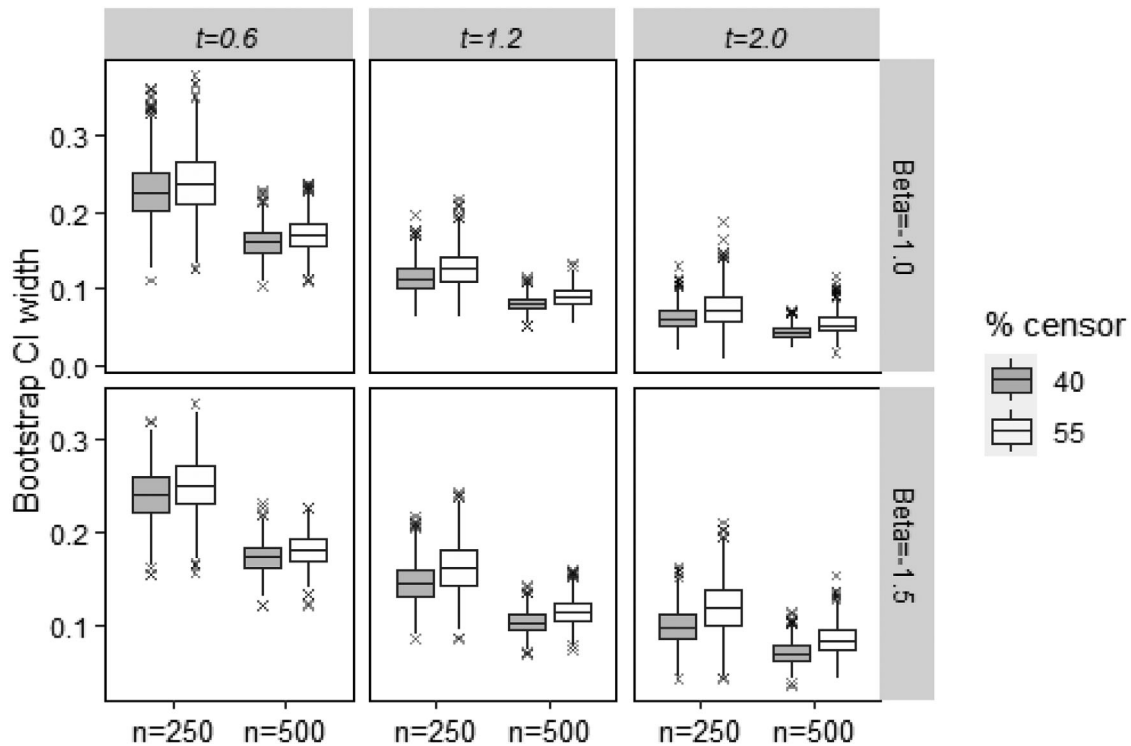


FIGURE 4 Boxplots for the width of the estimated bootstrap confidence interval obtained using the proposed method.

TABLE 4 Estimated time-dependent area under the precision–recall curve and associated 95% bootstrap confidence interval using the proposed method and the Yuan approach for the risk scores, M_1 , M_2 and their difference, and the prediction times, $t = 3, 6$ years.

t	Marker	Proposed		Yuan	
		Estimate	95% CI	Estimate	95% CI
3	M_1	0.719	[0.604, 0.815]	0.726	[0.616, 0.823]
3	M_2	0.616	[0.485, 0.732]	0.621	[0.497, 0.738]
3	Difference	0.104	[-0.063, 0.272]	0.105	[-0.059, 0.266]
6	M_1	0.809	[0.729, 0.886]	0.814	[0.731, 0.890]
6	M_2	0.698	[0.609, 0.789]	0.713	[0.618, 0.798]
6	Difference	0.111	[-0.008, 0.230]	0.101	[-0.021, 0.222]

4.2 | Data analysis

In this example, using the precision–recall curve and the associated summary index, we evaluate the ability of both risk scores to predict the survival of patients. Figure 5 presents the estimated time-dependent precision–recall curve using the proposed approach and the method proposed by Yuan et al. (2018) for both risk scores, that is, M_1 and M_2 , at prediction time $t = 3$ years and $t = 6$ years. From the figure, both the proposed method and the Yuan approach resulted in very similar precision–recall curve estimates. This is true for both risk scores and prediction times.

Table 4 shows the estimated time-dependent area under precision–recall curve with 95% bootstrap confidence interval using the proposed method and the Yuan approach for the risk scores, M_1 , M_2 and their difference, and the prediction times, $t = 3, 6$ years. The bootstrap confidence intervals are computed using the percentile approach described in Section 2.3 from 2000 bootstrap samples with replacement. For the proposed method and at $t = 3$ years, the estimated $AUPRC_t$ of the risk score M_1 [$\widehat{AUPRC}_t = 0.719$; 95%CI : 0.604 – 0.815] is higher than M_2 [$\widehat{AUPRC}_t = 0.616$; 95%CI : 0.485 – 0.732]. This means, the risk score M_1 has a better classification ability compared to M_2 . This difference in classification ability between the two risk scores, however, is not statistically significant, as indicated by the estimated confidence interval of the difference, which includes the “null” value (i.e., 0). This is consistent with the results of the estimated time-dependent precision–recall curves given in Figure 5. Similar conclusions can be drawn from the estimated time-dependent

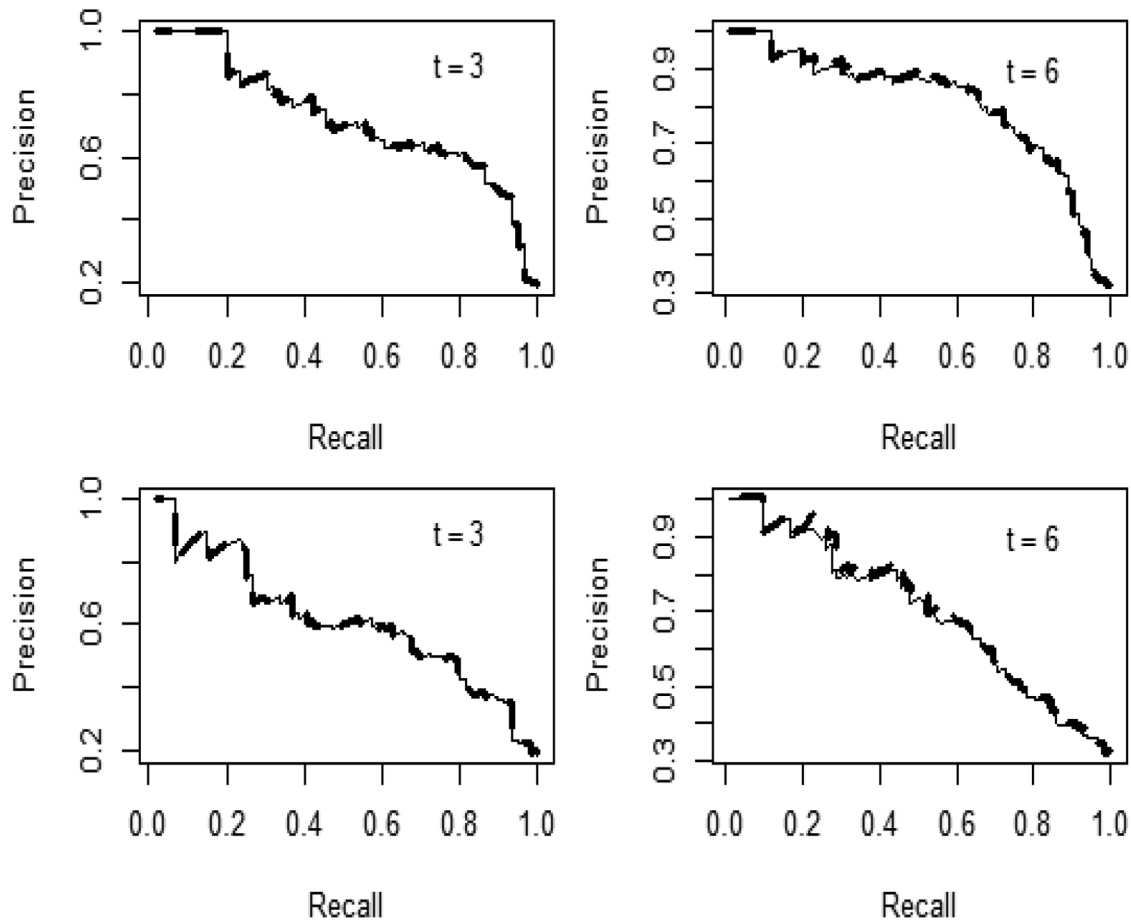


FIGURE 5 Time-dependent precision–recall curve estimates obtained using the proposed method (solid line) and the Yuan approach (dashed line) for M_1 (top row) and M_2 (bottom row) from Mayo PBC data with $t = 3$ and $t = 6$ years.

precision–recall curves and the associated area under the curves at $t = 6$ years. The estimated $AUPRC_t$ using the Yuan method for the risk score M_1 is larger than M_2 both for $t = 3$ years and $t = 6$ years; however, since zero is included in the estimated confidence interval of the difference, this difference is not statistically significant. Finally, in general, the estimated $AUPRC_t$ obtained using the proposed method is smaller than the estimate of Yuan approach.

5 | DISCUSSIONS

There is a growing use of prognostic scores in precision medicine for medical applications in order to identify subjects with a high risk of developing a particular condition. Nevertheless, the discriminatory ability of prognostic scores determines the quality of decisions based on them. Therefore, before clinicians use a prognostic score in routine clinical practice, the performance of the score should be properly assessed. To this end, this paper introduced a novel nonparametric time-dependent precision–recall curve and the associated area under the curve for right-censored time-to-event data. In our approach, we propose to estimate the conditional survival probabilities for subjects with unknown event statuses using the well-known Beran estimator. We used the selection method proposed by Sheather and Jones (1991) to choose the smoothing parameter required for kernel weight calculation in the Beran estimator. In addition, a nonparametric bootstrap approach is suggested to estimate variability and make inferences about the time-dependent area under precision–recall curve. The R package `tdPRC`, which implements the proposed method, is available in the Supporting Information.

The finite sample performance of the proposed time-dependent area under precision–recall curve estimation method is assessed through a simulation study. The finding of the simulation study revealed good performance for the proposed estimator with negligible bias and MSE. Moreover, the performance of the estimator improves with increase in sample size. Compared to the method that exists in the literature, the proposed estimator has shown to have smaller MSE. An

interesting finding of the study is that the Yuan approach, which is based on the IPCW, resulted in estimated $AUPRC_t$ values that exceeded 1 when the independence assumption is violated. This is due to the fact that, under a violation of this assumption, the sum of the estimated weights will not be anymore than 1. In our simulation study, we also investigated performance of the nonparametric bootstrap-based variance and confidence interval estimation method. The result indicated that the estimation method performed well since the average SD obtained from the bootstrap estimate is very close to the ESD and the CPs are also close to the desired nominal level. Finally, to illustrate the practical use of the proposed method in real-world data, we applied it to the PBC data set. This data set contains two risk scores that are derived using the Cox proportional hazards model (Heagerty & Zheng, 2005). Our objective was to assess and compare the ability of these scores to identify subjects with high risk of dying due to primary biliary cirrhosis disease. The time-dependent precision–recall curve and the associated area under the curve is estimated at two different times: 3 and 6 years. From the results, both scores have similar and reasonable discriminatory ability. The estimated confidence interval of the difference of area under the precision–recall curve of the two biomarkers indicated no significant difference as the confidence interval include the null value.

Finally, the proposed method is developed for right-censored data and hence extending it to other types of censored data, for example interval-censored, would be an interesting future research topic.

ACKNOWLEDGMENTS

Ding-Geng Chen acknowledge support from the South Africa National Research Foundation (NRF) and South Africa Medical Research Council (SAMRC; South Africa DST-NRF-SAMRC SARChI Research Chair in Biostatistics, Grant Number 114613). Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF and SAMRC.


CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the Supporting Information of this paper.

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to computational complexity.

ORCID

Kassu Mehari Beyene  <https://orcid.org/0000-0002-2067-6054>

Ding-Geng Chen  <https://orcid.org/0000-0002-3199-8665>

REFERENCES

- Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley.
- Beyene, K. M., & El Ghouch, A. (2020). Smoothed time-dependent receiver operating characteristic curve for right censored survival data. *Statistics in Medicine*, 39, 3373–3396.
- Beyene, K. M., & El Ghouch, A. (2022). Time-dependent ROC curve estimation for interval-censored data. *Biometrical Journal*, 64, 1056–1074.
- Beyene, K. M., El Ghouch, A., & Oulhaj, A. (2019). On the validity of time-dependent AUC estimation in the presence of cure fraction. *Biometrical Journal*, 61, 1430–1447.
- Blanche, P., Dartigues, J.-F., & Jacqmin-Gadda, H. (2013a). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32, 5381–5397.
- Blanche, P., Dartigues, J.-F., & Jacqmin-Gadda, H. (2013b). Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55, 687–704.
- Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The binormal assumption on precision-recall curves. In J. E. Guerrero (Ed.), *2010 20th international conference on pattern recognition* (pp. 4263–4266). IEEE.

- Cai, H., Yuan, Y., Zhou, Q. M., & Li, B. (2018). *APtools: Average positive predictive values (AP) for binary outcomes and censored event times*. <https://CRAN.R-project.org/package=APtools>. R package version 6.8.8.
- Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In W. W. Cohen, & A. Moore (Eds.), *Proceedings of the 23rd international conference on machine learning* (pp. 233–240). Association for Computing Machinery, New York, NY, United States.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7, 1–26.
- Etzioni, R., Pepe, M., Longton, G., Hu, C., & Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Medical Decision Making*, 19, 242–251.
- Fleming, T. R., & Harrington, D. P. (2011). *Counting processes and survival analysis*. John Wiley & Sons.
- Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2), 337–344.
- Heagerty, P. J., & packaging by Paramita Saha-Chaudhuri. (2022). *Survival ROC: Time-dependent ROC curve estimation from censored survival data*. <https://CRAN.R-project.org/package=survivalROC>. R package version 1.0.3.1.
- Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61, 92–105.
- Howe, C. J., Cole, S. R., Chmiel, J. S., & Munoz, A. (2011). Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias. *American Journal of Epidemiology*, 173, 569–577.
- Lambert, J., & Chevret, S. (2016). Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Statistical Methods in Medical Research*, 25, 2088–2102.
- Li, L., Greene, T., & Hu, B. (2018). A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Statistical Methods in Medical Research*, 27, 2264–2278.
- Martínez-Cambor, P., Bayón, G. F., & Pérez-Fernández, S. (2016). Cumulative/dynamic ROC curve estimation. *Journal of Statistical Computation and Simulation*, 86, 3582–3594.
- Martínez-Cambor, P., & Pardo-Fernández, J. C. (2018). Smooth time-dependent receiver operating characteristic curve estimators. *Statistical Methods in Medical Research*, 27, 651–674.
- Mookerjee, A., Meenakshi, R., & Hariprasad, R. (2007). Piepoc: A new prognostic index for advanced epithelial ovarian cancer—Japan multinational trial organisation oc01-01. *Indian Journal of Medical and Paediatric Oncology*, 28, 24–26.
- Moons, K. G., Royston, P., Vergouwe, Y., Grobbee, D. E., & Altman, D. G. (2009). Prognosis and prognostic research: What, why, and how? *BMJ*, 338, 1317–1320.
- Nadaraya, E. A. (1964). Some new estimates for distribution functions. *Theory of Probability & Its Applications*, 9, 497–500.
- Ozenne, B., Subtil, F., & Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, 68, 855–859.
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, 10, e0118432.
- Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53, 683–690.
- Slate, E. H., & Turnbull, B. W. (2000). Statistical models for longitudinal biomarkers of disease onset. *Statistics in Medicine*, 19(4), 617–637.
- Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10, 565–577.
- Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.
- Williams, C. K. (2021). The effect of class imbalance on precision-recall curves. *Neural Computation*, 33, 853–857.
- Yuan, Y., Zhou, Q. M., Li, B., Cai, H., Chow, E. J., & Armstrong, G. T. (2018). A threshold-free summary index of prediction accuracy for censored time to event data. *Statistics in Medicine*, 37, 1671–1681.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Beyene, K. M., Chen, D.-G., & Kifle, Y. G. (2024). A novel nonparametric time-dependent precision–recall curve estimator for right-censored survival data. *Biometrical Journal*, 66, 2300135. <https://doi.org/10.1002/bimj.202300135>