

A semiparametric method to measure predictive accuracy of covariates for doubly censored survival outcomes

Seungbong Han^{1,a}, JungBok Lee^b

^aDepartment of Applied Statistics, Gachon University, Korea;

^bDepartment of Clinical Epidemiology and Biostatistics, University of Ulsan College of Medicine, Asan Medical Center, Korea

Abstract

In doubly-censored data, an originating event time and a terminating event time are interval-censored. In certain analyses of such data, a researcher might be interested in the elapsed time between the originating and terminating events as well as regression modeling with risk factors. Therefore, in this study, we introduce a model evaluation method to measure the predictive ability of a model based on negative predictive values. We use a semiparametric estimate of the predictive accuracy to provide a simple and flexible method for model evaluation of doubly-censored survival outcomes. Additionally, we used simulation studies and tested data from a prostate cancer trial to illustrate the practical advantages of our approach. We believe that this method could be widely used to build prediction models or nomograms.

Keywords: double censoring, model evaluation, predictive ability, prostate cancer, variable selection

1. Introduction

It is important to evaluate the predictive accuracy of prognostic variables in order to build parsimonious but trustworthy prediction models. Our motivating example comes from a randomized clinical trial conducted to compare treatment arms (docetaxel with docetaxel vs. docetaxel only) (Attia *et al.*, 2008). This trial collected prostate cancer data from October 2002 to July 2005 in order to compare cancer progression rates and overall survival rates. In the present study, we were interested in assessing the predictive capacity of the treatment arms, baseline prostate-specific antigen (PSA), and other biomarker values such as hemoglobin and alkaline phosphatase levels, in relation to the time-to-death after the cancer relapses.

Multiple approaches can summarize the predictive accuracy of variables. For example, Heagerty *et al.* (2000) suggested a nonparametric estimator for sensitivity and specificity as well as constructed time-dependent receiver operating characteristic (ROC) curves based on a nonparametric kernel estimator. Heagerty and Zheng (2005) defined the sensitivities and specificities, termed ‘incident/static’ and ‘cumulative/dynamic’ respectively as well as estimated the time-dependent ROC using Cox models. Chambless and Diao (2006) introduced an AUC estimator using the Kaplan-Meier approach recursively, whereas Cai *et al.* (2006) estimated the time-dependent ROC via generalized linear models and proposed an asymptotic distribution theory for parameter estimates. The aforementioned procedures were proposed under the correctly specified working model, Uno *et al.* (2007) suggested

¹ Corresponding author: Department of Applied Statistics, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam 13120, Korea. E-mail: hanseungbong@gachon.ac.kr

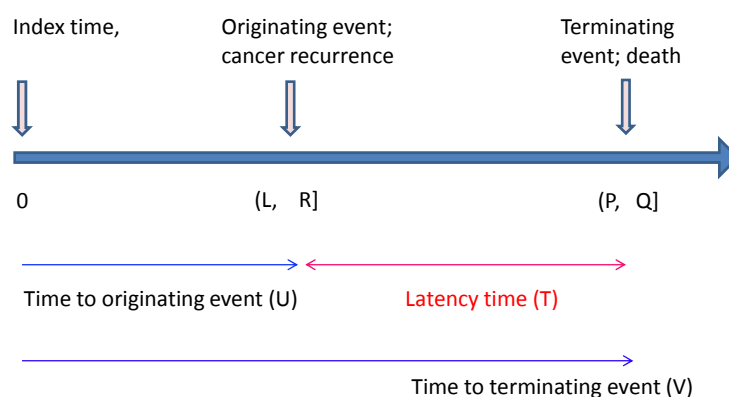


Figure 1: A schema for doubly-censored survival data. Originating event time U and the terminating event time V could be exact, right-censored, or interval-censored. The elapsed time is denoted by $T = V - U$. Both U and V are known only within an interval $(L, R]$ and $(P, Q]$. Here, $P = Q$ or $Q = \infty$ represent exact observations of death time or right-censored survival time.

consistent estimators for the sensitivity, specificity, and overall misclassification rate to predict future t -year survivors as well as removed the correct working model assumption. More recently, Song *et al.* (2012) proposed an overall predictive capacity across a certain time range without dependence on a Cox-type model. The majority of procedures mentioned so far function with right-censored survival data; however, the survivorship after recurrence of cancer is doubly-censored in the prostate cancer data from our motivating example.

Figure 1 shows a schema for doubly-censored data. The first event, the time to cancer recurrence, is assessed by periodic patient screening. In general, cancer recurrence time (U) is measured between the most recent relapse-free visit time (L) and the first visit time (R) at which the recurrence was detected. Furthermore, a secondary event, namely overall survival time V , is right-censored. Therefore, the doubly-censored data consists of two interval-censored endpoints $(L, R]$ and $(P, Q]$. Here, $P = Q$ or $Q = \infty$ represent exact observations of death time or right-censored survival time. Our interest lies in measuring the predictive accuracy of covariate combinations for the elapsed time between the two events ($T = V - U$).

To the best of our knowledge, a method for measuring the predictive accuracy of doubly-censored data has not yet been proposed. Recently, Chen *et al.* (2012) and Han *et al.* (2013) proposed the use of a negative predictive value (NPV) curve for right-censored or interval-censored survival data. They indicated that the NPV curve can be used to compare predictive accuracies of distinct covariate sets because it can be maximized at the covariate combination set, which is truly related to event times. We extend their approach towards doubly-censored survival data. One might consider a simple imputation approach to impute the cancer recurrence time by the mid-point of the interval and then apply survival techniques for right-censored data. However, this approach produces biased estimates for the survival function, coverage probabilities, confidence intervals and hazard ratio (Law and Brookmeyer, 1992; Zhang *et al.*, 2009). Therefore, the simple imputation approach may lead to a biased NPV curve and also result in an inappropriate predictive accuracy measurement. Several estimation methods have been proposed to model doubly-censored survival data (e.g., Cai and Cheng, 2004; Komárek *et al.*, 2005; Sun *et al.*, 2004; Yu, 2010). Recently, Han *et al.* (2014) proposed a flexible regression approach for modeling doubly-censored survival data. However, there are currently no

formal methods for model evaluation when building a prognostic model for doubly-censored survival outcomes.

Our proposed method uses a working model to obtain coefficient estimates for a linear combination; subsequently, we compute the predictive accuracy for a set of covariates semiparametrically. Hence, the predictive accuracy of the model can be compared without making specific model assumptions. Our simulation studies indicate that the NPV-based approach for doubly-censored survival data correctly identifies covariate combinations in various simulation settings. As detailed in the next section, semiparametric estimation of predictive accuracy provides a simple and flexible approach for model evaluation related to doubly-censored survival outcomes. The implementation effort is minimal, which is particularly advantageous when considering the relative scarcity of statistical software for doubly-censored survival settings. The remainder of this article is organized as follows: Section 2 describes the NPV-based model evaluation strategy, Section 3 uses extensive simulation results to illustrate the performance of our method with moderate sample sizes, Section 4 uses the method applied to the prostate cancer trial data, and Section 5 provides the concluding remarks.

2. A measure of predictive accuracy for doubly-censored data

2.1. Estimating the distribution function of the latency time

For the sake of simple notation, let $(L, R]$ and $(P, Q]$ be intervals of the originating event time U and the terminating event time V . That is, L and R are two consecutive visit times when the originating event such as cancer recurrences occurs and P and Q are also two consecutive visit times when the terminating event occurs. However, if the terminating event does not occur or right censored, the right end point Q would be ∞ (Sun, 2003). The observed doubly-censored data have the form $\{(L_i, R_i], (P_i, Q_i], Z_i; i = 1, \dots, n\}$, where Z_i is the p -dimensional covariate vector of subject i . The aim of conducting the analysis is to investigate the relationship between latency time ($T_i = V_i - U_i$) and the covariates Z_i .

In this study, we focus solely on subjects who experience the originating event; those subjects who do not experience the originating event are excluded from the sample. We also assume that a larger value of Z is associated with a high risk. Several methods are available in the literature to estimate the distribution function for the elapsed time $\Pr(T \leq t)$. First, in the seminal paper of De Gruttola and Lagakos (1989), they considered a nonparametric estimator based on a self-consistency equation. In contrast, Gómez and Lagakos (1994) proposed a two-step algorithm method to replace doubly-censored data with two separate sets of interval-censored data to resolve non-unique solutions due to a local maximum problem. It has been known that their method may suffer from a loss of efficiency. Later, Sun (1997) introduced a conditional likelihood approach to estimate the cumulative distribution function of T and U . His method handles the truncation issue as well as double censoring. The conditional likelihood method plays a central role to measure the predictive accuracy and we briefly explain as follows.

First, assume T and U takes only discrete values. We denote the cumulative distribution and the corresponding density function of T by $F(t)$ and $f(t)$. In parallel, $G(t)$ and $g(t)$ denote the cumulative distribution and the density function of U . Suppose the infection time U is truncated on the interval $(A, B]$ and $(A, B] = (0, \infty]$ means doubly censored failure time data. For each patient, the observed data would be $\{(L_i, R_i], (P_i, Q_i], (A_i, B_i]; i = 1, \dots, n\}$. Possible point mass points for U and T are denoted by $\gamma_0 = 0 < \gamma_1 < \dots < \gamma_r$ and $w_0 = 0 < w_1 < \dots < w_m$. Then $\gamma_0 = 0 < \gamma_1 < \dots < \gamma_r$ would be the ordered distinct elements of $\{0, L_i, L_i + 1, \dots, R_i; i = 1, \dots, n\}$ and $w_0 = 0 < w_1 < \dots < w_m$ would be the ordered distinct elements of $\{0, P_i - R_i, P_i - R_i + 1, \dots, Q_i - L_i; i = 1, \dots, n\}$. Note that

$g_j = G(\gamma_j) - G(\gamma_{j-1})$ and $f_k = F(w_k) - F(w_{k-1})$. Given $U_i \in (L_i, R_i]$, the derived conditional likelihood (LC) is

$$\prod_{i=1}^n \frac{\sum_{u=L_i}^{R_i} g(u)[F(U_i - u) - F(L_i - u - 1)]}{\sum_{u=L_i}^{R_i} g(u)[F(B_i - u) - F(A_i - u - 1)]}. \tag{2.1}$$

If the f 's are known, the problem would be reduced to maximize LC with respect to g 's, subject to the constraint $\sum_{j=1}^r g_j = 1$ and $\sum_{k=1}^m f_k = 1$. Let us define an indicator function ψ_{ij} of the event $\gamma_j \in (L_i, R_i]$. Also let $\psi_{ij}^\dagger = \sum_{u=P_i-\gamma_j}^{Q_i-\gamma_j} f_u$ if $\gamma_j \in (L_i, R_i]$ or $\psi_{ij}^\dagger = 1$, otherwise. With respect to the truncation interval, let $\varphi_{ij} = \sum_{u=A_i-\gamma_j}^{B_i-\gamma_j} f_u$ if $\gamma_j \in (L_i, R_i]$ or $\varphi_{ij} = 1$, otherwise. Then the LC is

$$\prod_{i=1}^n \frac{\sum_{j=1}^r g_j \psi_{ij} \psi_{ij}^\dagger}{\sum_{j=1}^r g_j \psi_{ij} \varphi_{ij}}. \tag{2.2}$$

Here, the LC is a function of the $\{g_j\}$ and can be similarly written as a function of the $\{f_k\}$. Based on two self-consistency equations, the conditional likelihood method attains superior efficiency to estimate G and F (Sun, 1997).

2.2. Semiparametric evaluation of the predictive accuracy

Suppose Z is a p -dimensional covariate vector and $\beta^T Z$ is a linear combination of Z under a regression model. We assume a semiparametric linear transformation model which means that the conditional survival function given Z , $\Pr(T > t|Z) = g(t, \beta^T Z)$ where $g(t, \beta^T Z)$ is a strictly decreasing and differentiable function for $\beta^T Z$ for all $t > 0$. For a pre-chosen fixed time t , Han *et al.* (2014) defined a time-dependent negative predictive function NPV(t, ν) as

$$\text{NPV}_{Z,\beta}(t, \nu) = \Pr\{T > t | F_{\beta^T Z}(\beta^T z) \leq \nu\}, \tag{2.3}$$

where β and $F_{\beta^T Z}$ represent regression coefficient effects and the cumulative distribution function of $\beta^T Z$, respectively. Similarly, a time-dependent positive predictive function PPV(t, ν) can be defined as

$$\text{PPV}_{Z,\beta}(t, \nu) = \Pr\{T < t | F_{\beta^T Z}(\beta^T z) > \nu\}. \tag{2.4}$$

One can easily show PPV(t, ν) is a monotone increasing function of NPV(t, ν) for the fixed values of t and ν ($0 \leq \nu \leq 1$). Han *et al.* (2014) and Chen *et al.* (2012) showed that the NPV function achieves its maximum at the true linear combination of covariates; therefore, the maximization is established uniformly for all values of t and ν among possible combinations of covariates. The predictive accuracy of different covariate combinations can also be compared. In certain analyses, one may be interested in overall predictive accuracy regardless of the pre-chosen time t . Following Han *et al.* (2014), an area under the NPV curve and an integrated summary measure can be derived as

$$\text{NAUC}_\beta(t) = \int_0^1 \text{NPV}_{Z,\beta}(t, \nu) d\nu \tag{2.5}$$

and

$$I_\beta(\tau) = \int_0^\tau \text{NAUC}_\beta(t) \times w(t) dt, \tag{2.6}$$

where $w(t)$ is a certain weight function. If $w(t)=1$, then $I_\beta(\tau)$ is an average of $\text{NAUC}_\beta(t)$ over $[0, \tau]$. Here τ can be chosen as a maximum event time support. The natural estimator of $\text{NPV}_{Z,\beta}(t, \nu)$ for doubly-censored data is

$$\widehat{\text{NPV}}_{Z,\hat{\beta}}(t, \nu) = \widehat{\Pr}\{T > t | F_n(\hat{\beta}^T z) \leq \nu\}, \quad (2.7)$$

where $\hat{\beta}$ is an estimated regression coefficient vector.

In (2.7), $\widehat{\text{NPV}}_{Z,\hat{\beta}}(t, \nu)$ is a nonparametric maximum likelihood estimate (NPMLE) of the marginal elapsed time function in the subgroup with $F_n(\hat{\beta}^T z) \leq \nu$, where $F_n(\cdot)$ is the estimated cumulative distribution (Sun, 1997). Even if a truncation is not present, it does not affect the estimation of the marginal elapsed time function f although the LC is different from the full likelihood. There are several ways to estimate β ; however, we subsequently use a method proposed by Han *et al.* (2013), which flexibly accommodates various survival models for doubly-censored data. The NPMLE based on the conditional likelihood method is consistent; therefore, $I_\beta(\tau)$ and $\text{NAUC}_\beta(t)$ are also consistently estimated by substituting $\widehat{\text{NPV}}_{Z,\hat{\beta}}(t, \nu)$ based on the continuous mapping theorem.

3. Simulation studies

Simulation studies were devised to evaluate the performance of the proposed method. When conducting these studies, numerical data are first generated within two different frameworks: a proportional hazards (PH) model ($\lambda(t|Z) = \lambda_0(t) \exp\{\beta^T Z\}$) and a log-linear (LL) model ($\log T = \beta^T Z + W$). Under each framework, samples with sizes $n = 50, 100, 200$, and 500 are considered, and a latency time T is generated based on covariates $Z = (Z_1, Z_2, Z_3, Z_4)$. Covariates of Z_1, Z_2, Z_3 , and Z_4 are generated independently: $Z_1 \sim N(1, 0.5)$, $Z_2 \sim \text{Bernoulli}(0.5)$, $Z_3 \sim \text{Unif}(0, 2)$ (PH) or $Z_3 \sim \text{Unif}(-2, 0)$ (LL), and $Z_4 \sim N(2, 0.5)$ (PH) or $Z_4 \sim \text{Unif}\{1, 2, 3, 4\}$ (LL). Vectors of regression parameters are fixed as $\beta = (-1, -1, -1, -1)$ (PH) or $(0.5, 0.5, 0.5, 0.5)$ (LL). We also assume that $\lambda_0(t) = t$ and $\exp(W)$ follows the Gompertz(1, 1) distribution. After generating the originating event time U from a PH or LL model, the terminating event time V is created by adding T and U . The originating and terminating event times are interval-censored between two consecutive visits. A possible inter-visit time process is drawn from a uniform $\text{Unif}(1, 4)$ distribution. The average widths of the censored intervals were 3.27 and the resulting terminating event times were subject to around 10% random-right censoring.

We compare five different regression models: (M1) $T \sim Z_1 + Z_2 + Z_3 + Z_4$, (M2) $T \sim Z_1 + Z_2 + Z_3$, (M3) $T \sim Z_1 + Z_2$, (M4) $T \sim Z_1 + Z_4 + X_2 + X_3$, and (M5) $T \sim Z_1 + Z_2 + Z_3 + Z_4 + X_1$. The first model is the correct model, whereas important prognostic factors are ignored in the second and third models. In the fourth and fifth models, we generate and include the noise variables $X_1 \sim \text{Unif}(-2, -1)$, $X_2 \sim N(2, 0.5)$, and $X_3 \sim N(-2, 1)$. To estimate NPV, $\text{NAUC}(t)$, and $I(\tau)$, we need a β estimate; for this, we use a regression method proposed by Han *et al.* (2013). To ensure that the time points for $\widehat{\text{NAUC}}_\beta(t)$ and τ are fixed across simulation-runs, we generate 100,000 event times and select the first (t_1), second (t_2), and third (t_3) quartiles for $\widehat{\text{NAUC}}_\beta(t)$ and its maximum for $\hat{I}_\beta(\tau)$. Tables 1 and 2 show simulation results based on 1,000 simulation-runs for the PH and LL models, respectively.

Based on the simulation results, the correct model, M1, produces the highest $\widehat{\text{NAUC}}_\beta(t)$ and $\hat{I}_\beta(\tau)$ values compared to those from M2, M3 and M4. However, M5 provides similar results to M1. Therefore, M5, a model that includes noise variables, apparently continues to have good predictive ability. Similar simulation results have previously been reported, for example, by Steyerberg *et al.* (2001) and Steyerberg (2009). In summary, the omission of a true predictor from a model appears to severely impact its predictive ability compared to a model that includes noise. Additionally, the inclusion of

Table 1: In the proportional hazards model, estimates for NAUC(t) and $I(\tau = 145)$ values are compared for $n = 50, 100, 200,$ and 500

n	t	M1	M2	M3	M4	M5
50	t_1	0.804	0.787	0.762	0.773	0.804
	t_2	0.659	0.625	0.586	0.598	0.652
	t_3	0.429	0.398	0.359	0.369	0.424
	$\hat{I}_{\beta}(\tau)$	18.287	17.458	16.400	16.648	18.121
100	t_1	0.812	0.784	0.759	0.767	0.801
	t_2	0.649	0.626	0.587	0.591	0.651
	t_3	0.418	0.395	0.354	0.360	0.419
	$\hat{I}_{\beta}(\tau)$	17.983	17.263	16.191	16.385	17.937
200	t_1	0.795	0.784	0.762	0.763	0.796
	t_2	0.647	0.625	0.589	0.589	0.648
	t_3	0.418	0.396	0.358	0.359	0.418
	$\hat{I}_{\beta}(\tau)$	18.291	17.479	16.464	16.524	18.108
500	t_1	0.796	0.783	0.762	0.763	0.796
	t_2	0.648	0.625	0.590	0.588	0.648
	t_3	0.422	0.399	0.362	0.361	0.422
	$\hat{I}_{\beta}(\tau)$	8.275	17.548	16.536	16.554	18.180

Three time points (t) and τ are selected at the first, second, and third quartiles and at a maximum follow-up time based on 100,000 event times. Results are based on 1,000 simulations.

Table 2: In the log-linear model, estimates for NAUC(t) and $I(\tau = 54)$ values are compared for $n = 50, 100, 200,$ and 500

n	t	M1	M2	M3	M4	M5
50	t_1	0.878	0.867	0.861	0.874	0.879
	t_2	0.687	0.634	0.606	0.656	0.672
	t_3	0.305	0.261	0.236	0.275	0.300
	$\hat{I}_{\beta}(\tau)$	4.778	4.384	4.055	4.443	4.728
100	t_1	0.876	0.870	0.868	0.874	0.877
	t_2	0.679	0.634	0.618	0.665	0.681
	t_3	0.298	0.243	0.222	0.274	0.297
	$\hat{I}_{\beta}(\tau)$	4.788	4.144	3.930	4.362	4.583
200	t_1	0.872	0.869	0.864	0.870	0.873
	t_2	0.679	0.634	0.618	0.662	0.680
	t_3	0.312	0.254	0.234	0.287	0.312
	$\hat{I}_{\beta}(\tau)$	4.784	4.241	4.037	4.474	4.681
500	t_1	0.872	0.868	0.864	0.870	0.872
	t_2	0.677	0.637	0.620	0.664	0.677
	t_3	0.316	0.255	0.236	0.292	0.317
	$\hat{I}_{\beta}(\tau)$	4.726	4.183	3.994	4.418	4.626

Three time points (t) and τ are selected at the first, second, and third quartiles and at a maximum follow-up time based on 100,000 event times. Results are based on 1,000 simulations.

noise variables has a small effect on the accuracy of the model.

4. The randomized phase II prostate cancer trial example

Attia *et al.* (2008) developed a database of 67 prostate cancer patients with data collected from October 2002 to July 2005. In this dataset, the time-to-cancer progression and the time-to-death were recorded, along with covariates of the treatment arms (docetaxel with doxercalciferol vs. docetaxel only), baseline values for PSA levels, the Eastern Cooperative Oncology Group (ECOG) performance

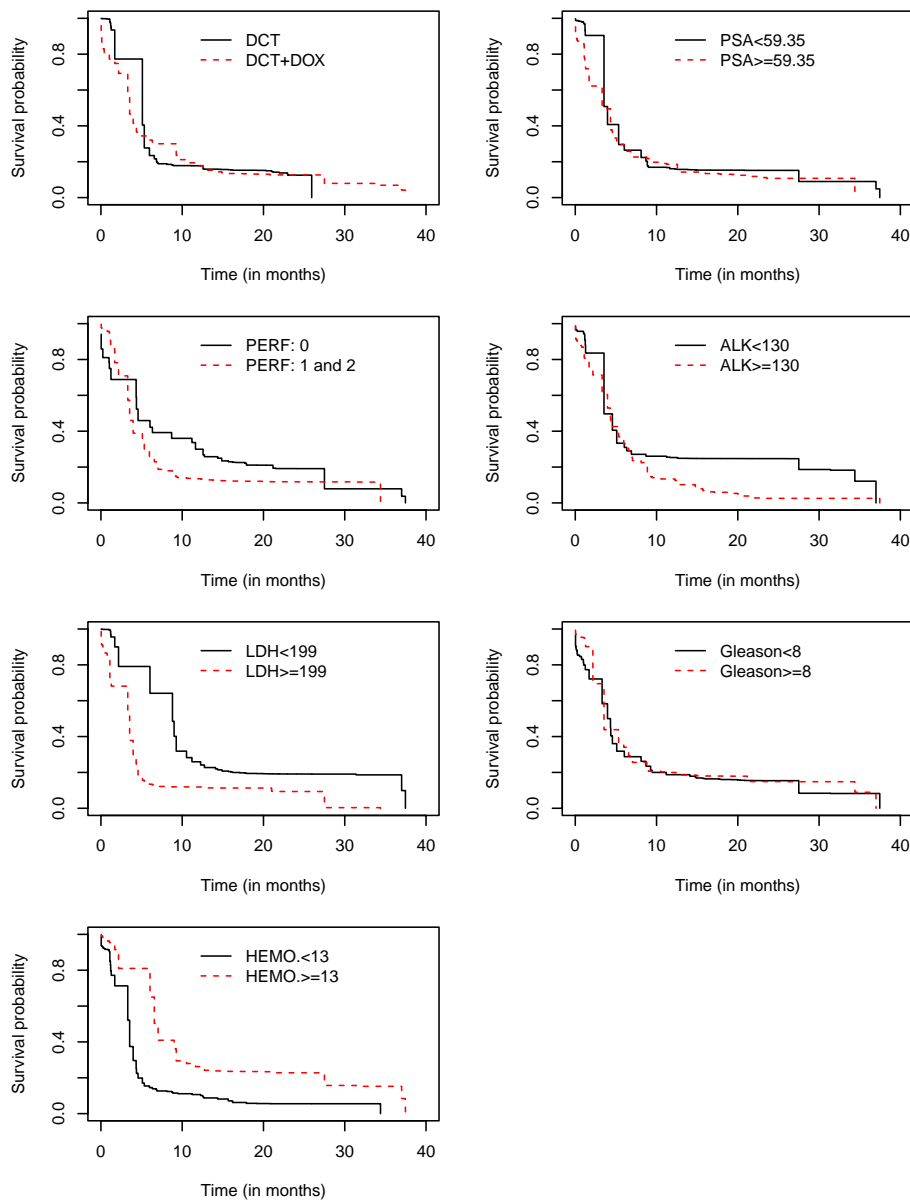


Figure 2: Prostate cancer trial example: assessing the proportional hazards assumption. Transformed survival curves vs. time in months are shown. DCT = docetaxel; DOX = doxorubicin; PSA = prostate-specific antigen; PERF = ECOG performance status; ALK = alkaline phosphatase; LDH = lactate dehydrogenase; HEMO. = hemoglobin.

status, age, hemoglobin levels, Gleason scores, and levels of lactate dehydrogenase and alkaline phosphatase. This study was designed to compare PSA responses between the two treatment arms in prostate cancer patients. As secondary endpoints, disease-free survival (DFS) and overall survival

Table 3: Prostate cancer trial example: $\widehat{\text{NAUC}}_{\beta}(t)$ and $\hat{I}_{\beta}(\tau)$ are estimated for $\tau = 40$ months for each covariate

Variable	$\widehat{\text{NAUC}}_{\beta}(t)$			$\hat{I}_{\beta}(\tau)$
	12	24	30	
ARM	0.160	0.113	0.000	6.997
PSA	0.160	0.144	0.085	8.110
PERF	0.139	0.117	0.103	7.173
ALK	0.266	0.243	0.138	10.604
LDH	0.231	0.141	0.133	10.163
GLEASON	0.141	0.092	0.056	7.274
HEMO	0.243	0.205	0.137	10.551

Abbreviations are defined in Figure 2.

Table 4: Prostate cancer trial example: $\widehat{\text{NAUC}}_{\beta}(t)$ and $\hat{I}_{\beta}(\tau)$ are estimated for several models

Model	$\widehat{\text{NAUC}}_{\beta}(t)$			$\hat{I}_{\beta}(\tau)$
	12	24	30	
ALK+HEMO	0.272	0.234	0.143	11.066
ALK+HEMO+LDH	0.284	0.248	0.177	11.304
ALK+HEMO+LDH+PSA	0.251	0.224	0.161	10.936
ALK+HEMO+LDH+PSA+GLEASON	0.217	0.195	0.135	9.986
ALK+HEMO+LDH+PSA+GLEASON+PERF	0.184	0.166	0.126	8.906
ALK+HEMO+LDH+PSA+GLEASON+PERF+ARM	0.179	0.142	0.076	8.429

Abbreviations are defined in Figure 2.

(OS) times were of interest, and Attia *et al.* (2008) performed two analyses for DFS and OS times. We use this dataset to identify risk factors related to mortality after cancer progression. In their study, Attia *et al.* (2008) measured disease progression by serial PSA scores; therefore, progression was regarded as occurring between the last progression-free visit and the first progression detection visit. Thus, we can see that time-to-cancer progression is interval-censored. Additionally, survival after the cancer progression is either right-censored or interval-censored.

Figure 2 illustrates the estimated survival probability according to covariate levels that use the conditional likelihood method of Sun (1997). The cut-off values for the continuous covariates were the median values. Results suggested that the added doxercalciferol does not alter the survival rate after cancer progression. In addition, baseline PSA and the Gleason score were not associated with survivorship after the progression. However, high lactate dehydrogenase and alkaline phosphatase levels or low hemoglobin levels were associated with an increased risk of mortality after the cancer progressed. Table 3 shows the estimated $\widehat{\text{NAUC}}_{\beta}(t)$ and $\hat{I}_{\beta}(\tau)$ for each covariate, with NAUC measured at 12, 24, and 30 months and the maximum follow-up time $\tau = 40$. The coefficient β is estimated under the proportional hazards model assumption, which is a reasonable assumption because the survival curves are parallel (Figure 2). Alkaline phosphatase produced the highest $\widehat{\text{NAUC}}_{\beta}(t)$ and $\hat{I}_{\beta}(\tau)$, and hemoglobin and lactate dehydrogenase were important prognostic factors. After ranking the prognostic factors by $\hat{I}_{\beta}(\tau)$, we also considered possible covariate combinations and obtained the $\widehat{\text{NAUC}}_{\beta}(t)$ and $\hat{I}_{\beta}(\tau)$. Table 4 shows that the combination of alkaline phosphatase, hemoglobin, and lactate dehydrogenase levels provides the best predictive ability for survivorship following cancer progression. It also appears that adding more covariates does not increase the power of this predictive ability.

We also fit a multivariable regression model with the three selected covariates (Table 5). High lactate dehydrogenase and alkaline phosphatase levels or low hemoglobin levels appear to increase

Table 5: Prostate cancer trial example: $NAUC_{\beta}(t)$ and $I_{\beta}(\tau)$ are estimated

Covariate	COEF	HR	95% CI		<i>p</i> -value
			LB	UB	
ALK	0.332	1.394	0.776	2.501	0.266
HEMO	-0.187	0.829	0.613	1.122	0.225
LDH	0.208	1.231	0.506	2.995	0.647

COEF = estimated regression coefficient; HR = hazards ratio; LB = lower bound; UB = upper bound; all other abbreviations are defined in Figure 2.

the mortality risk after cancer progression; however, the result was not statistically significant for any variable. This might be because there is a lack of statistical power for the three variables in the doubly-censored data setting. However, even variables with a lack of statistical power could be selected if covariates are selected in terms of their proposed predictive accuracy. In conclusion, we found that the proposed method can be used to assess whether or not important predictive covariates are included in the model.

5. Discussion

In the last 5 years, over 20,000 papers have been published in the biomedical literature that include the terms ‘prediction model’ or ‘nomogram’ (PubMed search: May 28, 2015). This level of research perhaps exists because building a reliable prediction model is thought to be imperative for individualized medicine (i.e., ‘from pre-womb to tomb’). When building such a model, predictor selection is a primary task. It is not reasonable to exclude predictors with *p*-values higher than the marginal value of 5% because prediction is about estimation rather than hypothesis testing. Methods based on ROC curves have been widely used in biomedical studies to evaluate predictors or markers.

In this study, we introduced a semiparametric method to measure predictive accuracy for doubly-censored survival data. NPV-based $NAUC(t)$ and $I(\tau)$ values provide a useful framework to evaluate the prognostic and diagnostic power of prediction rules. Such a framework makes very few assumptions and requires only regression coefficient estimates based on a working model. The proposed method compares predictive accuracy measurements for several candidate models; however, it would also be necessary to consider additional aspects of the models, such as discrimination capabilities, calibration, and clinical usefulness, before selecting a final model.

The proposed method can be extended to the left truncated and doubly censored survival data because the conditional likelihood method produces the estimate of the cumulative distribution function for the latency time T under the left truncation and doubly censoring setting. Zheng *et al.* (2008) measured the predictive value based on the PPV for the right censored survival data. To estimate the PPV, they proposed two methods which are a nonparametric approach and a semiparametric approach. While the semiparametric approach uses the Cox proportional hazards regression model, a cumulative baseline hazard function needs to be estimated. The cumulative baseline hazard function can be estimated based on the Breslow estimator for right censored survival data. However, it is a hard task for the doubly-censored data because of the censoring structure and may require a Bayesian MCMC approach which can be employed for the NPV and the subsequent quantities.

Our goal is to build parsimonious but trustworthy prediction models. In general, building a prediction model depends on a large data size (for example, over 300). In that case, the small number of observation due to the sub-grouping may not be a serious problem. In fact, our proposed method works well under the small to moderate sample sizes in our extensive simulation studies. However, we would not like to recommend the usage of the proposed method under the sample size less than

20 as one of the reviewers pointed it out. Another advantage of the proposed method is that we do not use the proportionality assumption compared to Zheng *et al.* (2008). To allow users to implement the proposed model evaluation for doubly-censored survival outcomes, the required R program can be found at <http://hanseungbong.wordpress.com/>.

Acknowledgments

This work was supported by the Gachon University research fund of 2015 (GCU-2015-0065).

References

- Attia S, Eickhoff J, Wilding G, McNeel D, Blank J, Ahuja H, et al. (2008). Randomized, double-blinded phase II evaluation of docetaxel with or without doxercalciferol in patients with metastatic, androgen-independent prostate cancer, *Clinical Cancer Research*, **14**, 2437–2443.
- Cai T and Cheng S (2004). Semiparametric regression analysis for doubly censored data, *Biometrika*, **91**, 277–290.
- Cai T, Pepe MS, Zheng Y, Lumley T, and Jenny NS (2006). The sensitivity and specificity of markers for event times, *Biostatistics*, **7**, 182–197.
- Chambless LE and Diao G (2006). Estimation of time-dependent area under the ROC curve for long-term risk prediction, *Statistics in Medicine*, **25**, 3474–3486.
- Chen L, Lin DY, and Zeng D (2012). Predictive accuracy of covariates for event times, *Biometrika*, **99**, 615–630.
- De Gruttola V and Lagakos SW (1989). Analysis of doubly-censored survival data, with application to AIDS, *Biometrics*, **45**, 1–11.
- Gómez G and Lagakos SW (1994). Estimation of the infection time and latency distribution of AIDS with doubly censored data, *Biometrics*, **50**, 204–212.
- Han S, Andrei AC, and Tsui KW (2013). A flexible semiparametric modeling approach for doubly-censored data with an application to prostate cancer, *Statistical Methods in Medical Research*. Advanced online publication. <http://dx.doi.org/10.1177/0962280213498325>.
- Han S, Tsui KW, and Andrei AC (2014). Model evaluation based on the negative predictive value for interval-censored survival outcomes, *Statistical Methods in Medical Research*. Advanced online publication. <http://dx.doi.org/10.1177/0962280214554253>.
- Heagerty PJ, Lumley T, and Pepe MS (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker, *Biometrics*, **56**, 337–344.
- Heagerty PJ and Zheng Y (2005). Survival model predictive accuracy and ROC curves, *Biometrics*, **61**, 92–105.
- Komárek A, Lesaffre E, Härkänen T, Declercq D, and Virtanen JI (2005). A Bayesian analysis of multivariate doubly-interval-censored dental data, *Biostatistics*, **6**, 145–155.
- Law CG and Brookmeyer R (1992). Effects of mid-point imputation on the analysis of doubly censored data, *Statistics in Medicine*, **11**, 1569–1578.
- Song X, Zhou XH, and Ma S (2012). Nonparametric receiver operating characteristic-based evaluation for survival outcomes, *Statistics in Medicine*, **31**, 2660–2675.
- Steyerberg EW (2009). Selection of main effects. In *Clinical Prediction Models* (pp. 191–212), Springer, New York.
- Steyerberg EW, Eijkemans MJ, Harrell FE, and Habbema JDF (2001). Prognostic modeling with logistic regression analysis in search of a sensible strategy in small data sets, *Medical Decision Making*, **21**, 45–56.

- Sun J (1997). Self-consistency estimation of distributions based on truncated and doubly censored data with applications to AIDS cohort studies, *Lifetime Data Analysis*, **3**, 305–313.
- Sun J (2003). Statistical analysis of doubly interval-censored failure time data, *Handbook of Biostatistics*, **23**, 105–122.
- Sun L, Kim YJ, and Sun J (2004). Regression analysis of doubly censored failure time data using the additive hazards model, *Biometrics*, **60**, 637–643.
- Uno H, Cai T, Tian L, and Wei LJ (2007). Evaluating prediction rules for t -year survivors with censored regression models, *Journal of the American Statistical Association*, **102**, 527–537.
- Yu B (2010). A Bayesian MCMC approach to survival analysis with doubly-censored data, *Computational Statistics and Data Analysis*, **54**, 1921–1929.
- Zhang W, Zhang Y, Chaloner K, and Stapleton JT (2009). Imputation methods for doubly censored HIV data, *Journal of Statistical Computation and Simulation*, **79**, 1245–1257.
- Zheng Y, Cai T, Pepe MS, and Levy WC (2008). Time-dependent predictive values of prognostic biomarkers with failure time outcome, *Journal of the American Statistical Association*, **103**, 362–368.

Received March 17, 2016; Revised May 4, 2016; Accepted May 18, 2016