Threshold-free Measure for Assessing the Performance of Risk Prediction with Censored Data

by

Bingying Li

B.Sc. (Hons), North Carolina State University, 2011

Project Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2015

Approval

Name:	Bingying Li
Degree:	Master of Science (Statistics)
Title:	*Threshold-free Measure for Assessing the Performance of Risk Prediction with Censored Data*
Examining Committee:	**Dr. Tim Swartz** (chair) Professor

Dr. Qian Zhou
Supervisor
Assistant Professor

Dr. Yan Yuan
Co-Supervisor
Assistant Professor

Dr. Joan Hu
Internal Examiner
Professor

Date Defended:	24 July 2015

Abstract

The area under the receiver operating characteristic curve (AUC) is a popular threshold-free metric to retrospectively measure the discriminatory performance of medical tests. In risk prediction or medical screening, main interests often focus on accurately predicting the future risk of an event of interest or prospectively stratifying individuals into risk categories. Thus, AUC might not be optimal in assessing the predictive performance for such purposes. Alternative accuracy measures have been proposed, such as the positive predictive value (PPV). Yuan et al. [1] proposed a threshold-free metric, the average positive predictive value (AP), which is the area under the PPV versus true positive fraction (TPF) curve, when the outcome is binary disease status. In this thesis, we propose the time-dependent AP when the outcome is censored event time. Empirical estimates of the time-dependent AP (AP_{t_0}) are developed, where the inverse weighted probability technique is applied to deal with censoring. In addition, inference procedures — using bootstrap and perturbation resampling — are proposed to construct confidence intervals. We conduct simulation studies to investigate the performance of the proposed estimation and inference procedures in finite samples. The method is also illustrated through a real data analysis.

Keywords: ROC curve; precision-recall curve; AUC; AP; survival data; medical risk prediction model

Acknowledgements

I would like to express my sincerest gratitude to my supervising professor, Dr. Qian Michelle Zhou, for her constant guidance and support throughout my master's program at Simon Fraser University. My time at SFU would not have been the same if it were not for the opportunity to study, work, and research under her supervision. Leading by example, not only has she taught me to excel academically, I have also learned many valuable career skills.

I would also like to thank my thesis co-supervisor Dr. Yan Yuan. Even with a time-zone difference, her continued support helped make my thesis project possible.

Much gratitude is extended to the members of my thesis committee, Dr. Tim Swartz and Dr. Joan Hu. Their valuable time and indispensable input are my inspiration.

I am sincerely grateful for my professors who have spent much time to prepare me throughout my master's program, and the department staff who have helped me with every administrative matter. I am also thankful for my fellow students who have walked this journey together with me. The entire Department of Statistics and Actuarial Science has brought me much joy throughout my two years of study at SFU, with a great environment, and much academic and financial support.

I am forever grateful to God, my family, and fellow Christians, who have taught me to rejoice in hope, endure in tribulation, and persevere in prayer.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In medicine, appropriate treatment depends on accurate diagnosis of medical conditions. Usually, diagnostic tests include radiographic images, biochemical analysis of body fluids, etc. Considering medical diagnoses in a broader context, screening healthy subjects is also important since many diseases can be more successfully treated if they are detected earlier. For clinicians, the utility of risk prediction is determined by its ability to predict the risk of developing disease by a specific follow-up time point t_0 given information of the subject. That is, the aim is to predict $P(D_i = 1 \mid Z_i)$ where D_i indicates the i-th subject's disease status (i.e. $D_i = 1$ if the subject developed the disease, and $D_i = 0$, otherwise). Z_i is a marker which contains certain information of the subject i. Accurate risk prediction is an important component in disease prevention and treatment management. Several accuracy measures have been proposed in the following literature.

1.1 Literature Review

The most widely used accuracy measure is called the receiver operating characteristic (ROC) curve. The ROC curve first arose in the context of signal detection theory, which was developed in the 1950s and 1960s [2, 3] and has been popular in medical diagnostic research, especially in radiology [4]. The ROC curve displays the sensitivity and specificity of an ordinal or continuous marker for a binary disease status variable D. Details about the ROC curve are presented in Section 2.1. A parametric ROC regression model [5, 6] of a generalized linear model (GLM) form has been proposed to examine covariates that affect the discriminatory capacity of a biomarker, which can be expressed as

$$ROC_Z(u) = g\{\theta' Z + h_\alpha(u)\}, \;\; 0 \leq u \leq 1,$$

where u is the range of 1-specificity, Z denote covariates, g is the link function, and h is the baseline function specified for a parameter α. This model extended the classic binormal model for the ROC curve [7] to include covariates. Cai and Pepe [8] extended the aforementioned parametric ROC regression method of Pepe [5, 6] to a semiparametric form, which reduced the requirements for model specification and increased robustness. In addition, smooth nonparametric and semiparametric estimators of ROC curves were proposed by Zou et al. [9] and by Metz et al. [10], respectively.

Area under the ROC curve (AUC) is one of the most commonly used summary indexes of the ROC curve. AUC captures the inherent discriminatory capability of the diagnostic test. A larger AUC value corresponds to a better test performance. Partial area under the curve (pAUC) is another summary index when clinical interests lie only in a specific range of the sensitivity or specificity [11, 12].

In risk prediction, disease outcomes are time dependent, depending on factors such as the time to occurrence of the disease. In these situations, the aim is to predict the t_0-year risk, which is the probability $P(T \leq t_0 \mid Z)$ of developing the disease by a predetermined time t_0, where T is the time to occurrence of the disease. Heagerty et al. [13] proposed a time-dependent ROC curve with a single marker where the disease status $D(t) = I\left(T \leq t_0\right)$ is time-dependent, and $I\left(\cdot\right)$ is the indicator function. They proposed two ROC curve estimators that accommodate censored data. One is based on the Kaplan-Meier survival function method and the other is based on a nearest neighbor estimator for the bivariate distribution function. With multiple markers, the t_0-year risk can be estimated by Cox proportional hazards models [14] or linear transformation models [15]. Uno et al. [16] proposed a class of time-dependent GLMs which allows the effects of multiple covariates to vary with time. Based on the estimated risk, the time-dependent ROC curve can be estimated parametrically [17] or nonparametrically [16].

Aside from AUC, the positive predictive value (PPV) curve [18] was proposed to be useful in prospective accuracy evaluation. Zheng et al. [19] extended the PPV curve to censored data. In addition, Zheng et al. [20] also proposed a covariate-specific PPV curve based on semiparametric models with censored data. The precision-recall curve, which is the PPV versus true positive fraction (TPF) curve, is another metric which can be considered for risk prediction. This curve is widely used in the information retrieval community, where documents are assigned to two categories: ‘relevant’ and ‘not relevant’. In medical research, it is of great interest for clinicians to use a marker to identify high risk groups. People with marker values greater than a given cut-off value are classified as being in the high risk group, and they may be recommended for intensive follow-up or therapeutic treatment in order to improve long-term disease outcomes. The ability of the marker to identify high risk subpopulations may be better reflected by its PPV. Yuan et al. [1] proposed that the area under precision-recall curve, the average precision (AP), which is a threshold-free summary index of the precision-recall curve, could be an attractive alternative metric to AUC in detecting the high risk group. They have shown that AP could improve decision making for screening tests where a low prevalence disease rate is typical.

1.2 Objectives

In this thesis, we propose the time-dependent AP (AP_{t_0}) for censored time to event data in risk prediction. We derive its estimation and inference procedures through a nonparametrical approach. One of the challenges in the estimate derivation and inference procedures is that the event time may be subject to censoring, thus not all the event time are observable. The inverse probability weighting (IPW) technique [16] is used to deal with this issue. In addition, through simulation studies, we compare the two accuracy measures, time-dependent AUC and AP, when the distributions of the marker values differ among the subjects who develop the disease by time t_0 and those subjects who are disease-free by time t_0.

1.3 Outline

In Chapter 2, we derive the estimates of AUC and AP when the test outcomes are binary and censored event time, as well as the corresponding inference procedures via bootstrap and perturbation resampling. Chapter 3 presents simulation studies, where we investigate the finite sample performance of proposed estimates and inference procedures. In Chapter 4, we use a real-life dataset from the Childhood Cancer Survivor Study to examine the proposed time-dependent AUC and AP. In Chapter 5, we summarize this project and outline a few problems for future investigation.

Chapter 2

Method

In this chapter, we introduce several popular prediction accuracy measurements when the outcomes are either ‘binary disease status’ or ‘time to event’. A new metric to evaluate the predictive performance of risk prediction models is proposed, and the corresponding estimation and inference procedures are also provided.

2.1 Prediction accuracy measures with binary outcome

In epidemiology, the group of diseased individuals ($D = 1$) are called the cases, and the group of disease-free individuals ($D = 0$) are called the controls. Let Z denote a marker used in the diagnostic test, which takes ordinal or continuous values. It is usually assumed that higher values of the marker indicate higher risks of having the disease. With a threshold c, a positive test result is defined if $Z \geq c$. Otherwise, a negative test result is defined.

Several accuracy measures have been proposed to evaluate the performance of the diagnostic test. ROC curve is one of the most popular tools. An ROC curve is a plot of the TPF function versus false positive fraction (FPF) function for all possible thresholds c. The TPF is defined as the fraction of subjects with positive test results in the case group, and the FPF is defined as the fraction of subjects with positive test results in the control group. That is, $TPF(c) = P[Z > c \mid D = 1]$ and $FPF(c) = P[Z > c \mid D = 0]$. In the literature, the TPF is also called the sensitivity and $1-$FPF is also called the specificity. The ROC curve is the entire set of possible true and false positive fraction values with different threshold values c (i.e. $ROC(\cdot) = \{(FPF(c), TPF(c)),\ c \in (-\infty, +\infty)\}$). An ROC curve can also be written as $ROC(\cdot) = \{(u, ROC(u)),\ u \in (0, 1)\}$, where $ROC(u) = TPF\{FPF^{-1}(u)\}$.

AUC is one of the most commonly used threshold-free summary indexes of the ROC curve, it is defined as

$$AUC = \int_0^1 ROC(u)du = \int_{\infty}^{-\infty} TPF(c)dFPF(c).$$

It can be shown that AUC is the probability $AUC = P(Z_1 > Z_0)$ that the marker value of a randomly selected diseased subject is greater than that of a randomly selected disease-free subject, where Z_1 and Z_0 correspond to the marker values of a randomly chosen case and control, respectively. AUC captures the inherent discriminatory ability of the diagnostic tests. According to Yuan et al. [1], the range of AUC is [0.5, 1]. When a diagnostic (or screening) test is random, which is a useless test, AUC is equal to 0.5. Whereas when the diagnostic (or screening) test is perfect, which is regarded as the best test, AUC is equal to 1.

Besides the ROC curve and AUC, accuracy of the test can also be measured by looking at how well the test can predict the true disease status, such as the PPV and negative predictive value (NPV). The PPV is defined as the fraction of diseased subjects among the group with positive test results, and the NPV is defined as the fraction of non-diseased subjects among the group with negative test results. That is, $PPV(c) = P[D = 1 \mid Z > c]$ and $NPV(c) = P[D = 0 \mid Z \leq c]$. The precision-recall (PR) curve plots the PPV (precision) function versus the TPF (recall) function. That is, $PR(\cdot) = \{(TPF(c), PPV(c)),\, c \in (-\infty, +\infty)\}$. It can also be written as $PR(\cdot) = \{(u, h(u)),\, u \in (0, 1)\}$, where $h(u) = PPV\{TPF^{-1}(u)\}$.

Area under the precision-recall curve, which is also called the average precision (AP), was proposed by Yuan et al. [1]. It can be written as

$$AP = \int_0^1 h(u)du = \int_{\infty}^{-\infty} PPV(c)dTPF(c).$$

In Appendix A, we show that AP gives the expected probability that a randomly selected subject is diseased given his/her marker value is greater than that of a randomly selected diseased subject. That is, $AP = E[P(D_i = 1 \mid Z_i > Z_{1j})]$, where Z_{1j} stands for the marker value of a randomly selected diseased subject, and expectation is taken with respect to the distribution of Z_{1j}. Thus, AP captures the marker's ability to identify diseased individuals. According to Yuan et al. [1], the range of AP is [r, 1] where r represents the event rate. When a diagnostic (or screening) test is random, AP is equal to r. When the test is perfect, AP is equal to 1.

To estimate metrics TPF, FPF, PPV and NPV , we use an empirical nonparametric approach. Their estimates are

$$\widehat{TPF}(c) = \frac{\sum_{i=1}^n I(D_i = 1)I(Z_i > c)}{\sum_{i=1}^n (D_i = 1)},\ \widehat{FPF}(c) = \frac{\sum_{i=1}^n I(D_i = 0)I(Z_i > c)}{\sum_{i=1}^n (D_i = 0)},$$

$$\widehat{PPV}(c) = \frac{\sum_{i=1}^n I(D_i = 1)I(Z_i > c)}{\sum_{i=1}^n I(Z_i > c)},\ \widehat{NPV}(c) = \frac{\sum_{i=1}^n I(D_i = 0)I(Z_i \leq c)}{\sum_{i=1}^n I(Z_i \leq c)}.$$

Combining the aforementioned estimates, the estimated AUC and AP are

$$\widehat{AUC} = \int_{\infty}^{-\infty} \widehat{TPF}(c) d\widehat{FPF}(c) = \frac{\sum_{j=1}^{n}\sum_{i=1}^{n} I(D_j = 0) I(D_i = 1) I(Z_i > Z_j)}{\sum_{j=1}^{n} I(D_j = 0) \sum_{i=1}^{n} I(D_i = 1)},$$

$$\widehat{AP} = \int_{\infty}^{-\infty} P\hat{P}V(c) dT\hat{P}F(c) = \frac{1}{\sum_{j=1}^{n} I(D_j = 1)} \sum_{j=1}^{n} I(D_j = 1) \frac{\sum_{i=1}^{n} I(D_i = 1) I(Z_i > Z_j)}{\sum_{i=1}^{n} I(Z_i > Z_j)}.$$

2.2 Prediction accuracy measures with censored survival data

In some applications, the time to event (such as the occurrence of the disease) is of interest. The event time might be subject to censoring due to failure to follow up or end of study. Let C_i be the censoring time. With censoring, only (X_i, δ_i, Z_i) can be observed, where Z_i stands for the individual's marker, $X_i = \min(T_i, C_i)$, and $\delta_i = I(T_i \leq C_i)$. Given a prespecified time t_0, individuals with $T_i \leq t_0$ make up the case group; the individuals with $T_i > t_0$ constitute the control group. The time-dependent TPF, FPF, PPV and NPV are defined as $TPF_{t_0}(c) = P(Z > c \mid T \leq t_0)$, $FPF_{t_0}(c) = P(Z > c \mid T > t_0)$, $PPV_{t_0}(c) = P(T \leq t_0 \mid Z > c)$, and $NPV_{t_0}(c) = P(T > t_0 \mid Z \leq c)$.

The corresponding time-dependent AUC and AP are defined as

$$AUC_{t_0} = \int_{\infty}^{-\infty} TPF_{t_0}(c) dFPF_{t_0}(c) = P(Z_i > Z_j \mid T_i \leq t_0, T_j > t_0),$$

$$AP_{t_0} = \int_{\infty}^{-\infty} PPV_{t_0}(c) dTPF_{t_0}(c) = E[P(T_i \leq t_0 \mid Z_i > Z_j, T_j \leq t_0)].$$

Without censoring, AUC and AP can be estimated by the empirical estimates as

$$\widehat{AUC}_{t_0} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{n} I(T_j > t_0) I(T_i \leq t_0) I(Z_i > Z_j)}{\sum_{j=1}^{n} I(T_j > t_0) \sum_{i=1}^{n} I(T_i \leq t_0)},$$

$$\widehat{AP}_{t_0} = \frac{1}{\sum_{j=1}^{n} I(T_j \leq t_0)} \sum_{j=1}^{n} I(T_j \leq t_0) \frac{\sum_{i=1}^{n} I(T_i \leq t_0) I(Z_i > Z_j)}{\sum_{i=1}^{n} I(Z_i > Z_j)}.$$

With censoring, for subjects whose $X_i \leq t_0$ and $\delta_i = 0$, $I(T_i \leq t_0)$ are unknown, which can be regarded as missing values. Using the IPW technique by Uno et al. [16], the estimates of AUC_{t_0} and AP_{t_0} are given by

$$\widehat{AUC}_{t_0} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{n} I(X_j > t_0) \hat{w}_j I(X_i \leq t_0) \hat{w}_i I(Z_i > Z_j)}{\sum_{j=1}^{n} I(X_j > t_0) \hat{w}_j \sum_{i=1}^{n} I(X_i \leq t_0) \hat{w}_i},$$

$$\widehat{AP}_{t_0} = \frac{1}{\sum_{j=1}^{n} I(X_j \leq t_0) \hat{w}_j} \sum_{j=1}^{n} I(X_j \leq t_0) \hat{w}_j \frac{\sum_{i=1}^{n} I(X_i \leq t_0) \hat{w}_i I(Z_i > Z_j)}{\sum_{i=1}^{n} I(Z_i > Z_j)},$$

where $\hat{w}_i$ is the inverse of the probability that $I(T_i \leq t_0)$ is observed. More precisely, $\hat{w}_i = \frac{I(X_i \leq t_0)\delta_i}{\hat{G}(X_i)} + \frac{I(X_i > t_0)}{\hat{G}(t_0)}$, where $\hat{G}(\cdot)$ is the estimate of the survival function $G(\cdot)$ of censoring time C_i (i.e. $G(c) = P(C_i > c)$). If the censoring is independent of both the event time and the marker, then $G(\cdot)$ can be estimated by the Kaplan-Meier estimator. If the censoring depends on the marker, then Cox proportional hazard models could be fit to the censoring time depending on the marker and obtain the estimates of $G(\cdot \mid Z)$.

2.3 Inference

In this thesis, we focus on time-dependent AUC and AP. To construct confidence intervals, two methods can be applied. The first one is bootstrap, and the second one is perturbation resampling.

Bootstrap is a random sampling method with replacement; the procedure has no external input. Standard deviations and confidence intervals can be derived from numerous repetitions of bootstrapping with the same sample size n.

The perturbation resampling is also called the wild bootstrap [21, 22], which is widely used in survival analysis when the asymptotic variances are difficult to calculate. Let $\{V_i, i = 1, \ldots, n\}$ be n independent copies of a positive random variable V with mean 1 and variance 1. For example, V_i can be generated from an exponential distribution. The perturbed estimates of AUC and AP are obtained by

$$\widehat{AUC}^*_{t_0} = \frac{\sum_{j=1}^n \sum_{i=1}^n I(X_j > t_0)\hat{w}_j^* V_j I(X_i \leq t_0)\hat{w}_i^* V_i I(Z_i > Z_j)}{\sum_{j=1}^n I(X_j > t_0)\hat{w}_j^* V_j \sum_{i=1}^n I(X_i \leq t_0)\hat{w}_i^* V_i},$$

$$\widehat{AP}^*_{t_0} = \frac{1}{\sum_{j=1}^n I(X_j \leq t_0)\hat{w}_j^* V_j} \sum_{j=1}^n I(X_j \leq t_0)\hat{w}_j^* V_j \frac{\sum_{i=1}^n I(X_i \leq t_0)\hat{w}_i^* V_j I(Z_i > Z_j)}{\sum_{i=1}^n I(Z_i > Z_j) V_i},$$

where $\hat{w}_i^*$ is the perturbed $\hat{w}_i$, which can also be written as $\hat{w}_i^* = \frac{I(X_i \leq t_0)\delta_i}{\hat{G}^*(X_i)} + \frac{I(X_i > t_0)}{\hat{G}^*(t_0)}$. Here $\hat{G}^*(\cdot)$ is the perturbed estimate of $G(\cdot)$ with weights V_i. More precisely, $\hat{G}^*(t) = \exp\{-\hat{\Lambda}^*(t)\}$ and $\hat{\Lambda}^*(t) = \sum_{i=1}^n \frac{I(X_i \leq t)(1-\delta_i)V_i}{\sum_{k=1}^n I(X_k \geq X_i)}$ from the Nelson-Aalen estimator.

To construct the 95% confidence intervals for the AUC and AP using bootstrap or perturbation resampling, two approaches are applied. The first one is to find the 2.5th quantile from all of the resampling estimates as the lower limit, whereas the 97.5th quantile is the upper limit. The second approach is to apply the normal approximation. For example, the 95% confidence interval for AUC_{t_0} estimates can be calculated as

$$\begin{cases} \text{Lower Limit} = \widehat{AUC}_{t_0} - 1.96 \times \text{sd}(\widehat{AUC}^*_{t_0}) \\ \text{Upper Limit} = \widehat{AUC}_{t_0} + 1.96 \times \text{sd}(\widehat{AUC}^*_{t_0}) \end{cases},$$

where $\widehat{\mathrm{AUC}}_{t_0}$ is the point estimate, and $\mathrm{sd}(\widehat{\mathrm{AUC}}^*_{t_0})$ is obtained from bootstrap or perturbed resampling. This procedure is also applied to construct the confidence intervals of AP_{t_0}.

Chapter 3

Simulation Studies

To examine the finite sample performances of time-dependent AUC and AP estimators, we conducted a simulation study.

Let the event time $T_i = \exp(2+\epsilon_i)$, where ϵ_i is generated from a standard extreme value distribution. Given a prespecified time t_0, the marker values Z_i for the control group with $T_i > t_0$ are generated from a standard normal distribution; the marker values for the case group with $T_i \leq t_0$ are generated from a normal distribution $N(\mu_Z, \sigma_Z)$. In this simulation, we consider two sets of values for $\{\mu_Z, \sigma_Z\}$,which are $\{0.95, 1\}$ and $\{1.51, 2\}$. This setting will result in similar estimated values of time-dependent AUC but different time-dependent AP based on the studies of Yuan et al. [1]. The censoring time C_i is generated following $C_i = \min\{C_{1i}, C_{2i} + 1\}$, where $C_{i1} \sim \text{Uniform}(0, 40)$, and $C_{2i} \sim \text{Gamma}(4, 0.75)$. This configuration results in about 50% of censoring. With the realizations of T_i, Z_i, and C_i, let $X_i = \min(T_i,\, C_i)$ and $\delta_i = I\,(T_i \leq C_i)$.

In this simulation study, three prediction time points are considered so that they result in event rates $r = P(T_i \leq t_0)$ to be 0.01, 0.05 and 0.1. In addition, we generate the data $\{(X_i,\, Z_i,\, \delta_i),\, i = 1, \ldots, n\}$ with different sample sizes n to be 5000 and 10000. The following results are obtained based on 1000 repetitions. For each repetition, 1000 bootstrap resamples and 1000 perturbed resamples are generated to construct confidence intervals.

Table 3.1 and 3.2 below report the results for sample sizes 5000 and 10000, respectively. In each table, we show the following summary statistics: bias, empirical standard error (ESE), average standard errors from bootstrap (ASE^b), average standard errors from perturbation resampling (ASE^p), the empirical coverage probability from bootstrap (ECOVP^b) and the empirical coverage probability from perturbation resampling (ECOVP^p) using normal approximation. Specifically, the bias is calculated as the difference between the average of the point estimates of the 1000 replicates and the true value of the accuracy measures AUC_{t_0} or AP_{t_0}. In this simulation study, we obtain the true value via averaging the estimates of AUC_{t_0} and AP_{t_0} from a very large sample ($n = 100000$) over 10 repetitions without censoring. The ESE is equal to the empirical standard deviation of 1000 point estimates.

		TRUE	BIAS	ESE	ASE^b	ASE^p	ECOVP^b	ECOVP^p
r=0.01	AUC N(0.95,1)	74.99	0.04	2.99	2.95	2.91	94.4	93.9
	AUC N(1.51,2)	74.96	0.00	4.01	4.00	3.94	94.4	94.2
	AP N(0.95,1)	5.02	-0.09	1.50	1.36	1.31	85.7	85.3
	AP N(1.51,2)	31.15	-1.26	5.83	5.83	5.62	93.1	91.9
r=0.05	AUC N(0.95,1)	74.93	-0.04	1.56	1.53	1.52	94.2	94.3
	AUC N(1.51,2)	75.16	-0.12	2.09	2.03	2.02	94.3	94.5
	AP N(0.95,1)	16.72	-0.16	2.00	1.93	1.89	92.6	91.7
	AP N(1.51,2)	44.57	-0.69	3.25	3.10	3.08	93.4	93.2
r=0.1	AUC N(0.95,1)	74.92	-0.04	1.12	1.13	1.13	94	94.3
	AUC N(1.51,2)	75.02	0.04	1.46	1.46	1.46	95.1	95.4
	AP N(0.95,1)	28.22	-0.27	1.93	1.93	1.91	94.6	94.1
	AP N(1.51,2)	52.76	-0.20	2.20	2.20	2.19	95.2	95

Table 3.1: Result of Sample Size 5000.

*(*All the values shown have been multiplied by* 100*)*

		TRUE	BIAS	ESE	ASE^b	ASE^p	ECOVP^b	ECOVP^p
r=0.01	AUC N(0.95,1)	74.99	-0.06	2.00	2.09	2.07	95.9	95.5
	AUC N(1.51,2)	74.96	0.23	2.80	2.82	2.79	94.4	94.4
	AP N(0.95,1)	5.02	-0.05	1.06	1.02	1.00	90.3	89.8
	AP N(1.51,2)	31.15	-0.59	4.14	4.15	4.08	94.7	94
r=0.05	AUC N(0.95,1)	74.93	-0.02	1.04	1.08	1.08	95.7	95.5
	AUC N(1.51,2)	75.16	-0.07	1.45	1.43	1.43	94.3	94.4
	AP N(0.95,1)	16.72	0.00	1.42	1.40	1.38	93.6	92.9
	AP N(1.51,2)	44.57	-0.36	2.22	2.20	2.18	95.2	95.1
r=0.1	AUC N(0.95,1)	74.92	-0.03	0.77	0.80	0.80	96.5	96.5
	AUC N(1.51,2)	75.02	0.00	1.06	1.04	1.04	95.2	95
	AP N(0.95,1)	28.22	-0.11	1.37	1.39	1.37	95.7	95.3
	AP N(1.51,2)	52.76	-0.10	1.60	1.56	1.55	95	94.8

Table 3.2: Result of Sample Size 10000.

*(*All the values shown have been multiplied by* 100*)*

Based on the simulation results from finite samples, we can observe that

- The estimates have small bias in each scenario, and the bias decreases with the increase of sample size.

- The standard errors generated from bootstrap ASE^b and perturbation resampling ASE^p are close to the ESE. In addition, all of these standard errors decrease when the sample size increases.

- The empirical coverage probabilities from bootstrap and perturbation resampling have no substantial differences from each other, but perturbation resampling is more computation-friendly. Most of the coverage probabilities approach 95%. However, there are a few under-coverage situations for AP_{t_0}, which happens when both event rate and sample size are small. Therefore, a logit transformation is applied to improve the performance of time-dependent AP estimates. That is, $\widehat{new\_AP}_{t_0} = \log(\frac{\widehat{AP_{t_0}}}{1-\widehat{AP_{t_0}}})$. Table B.1 and B.2 in Appendix B have more details about the improved results on the coverage of the transformed estimator.

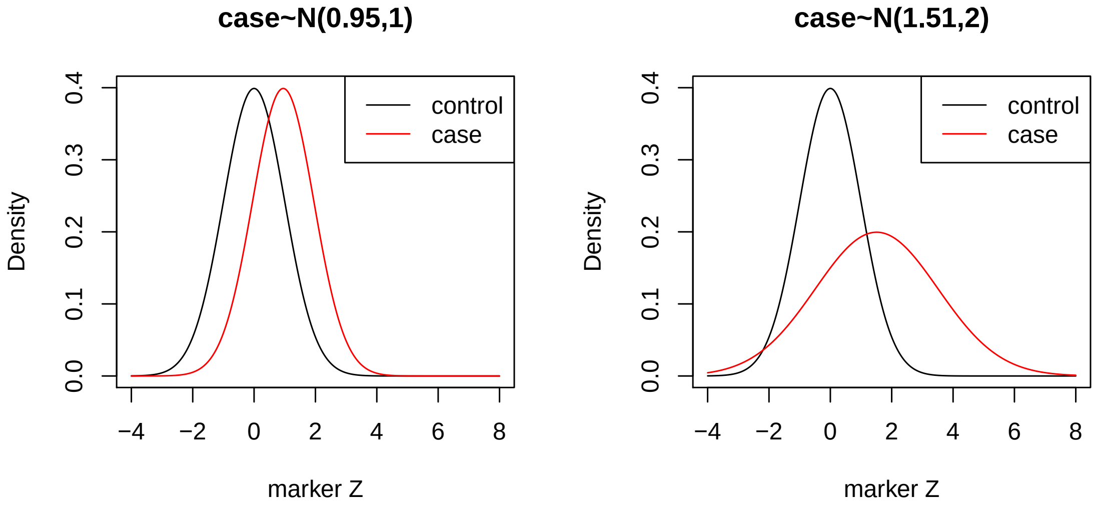

Figure 3.1: True Distribution Density Curves for Case and Control Marker Values ($r = 0.05$).

For different event rates ($r = 0.01, 0.05, 0.1$), the values of AUC_{t_0} are similar to each other and they are invariant to event rates. On the contrary, the values of AP_{t_0} are quite distinct from each other and they are sensitive to different event rates. In addition, we compare the time-dependent AUC and AP under these two scenarios where the distribution of the marker values for controls is the same (both $N(0,1)$) but the distribution of the marker values for cases are different ($N(0.95,1)$ versus $N(1.51,2)$). For example, when the event rate is 0.05, the AUC_{t_0}s are 0.749 and 0.751 for $N(0.95, 1)$ and $N(1.51, 2)$, and the AP_{t_0}s are 0.167 and 0.446, respectively.

Moreover, we consider that the two markers' difference in AP_{t_0} come from how different the distribution of the marker values among cases is from controls. Figure 3.1 shows the density curves of the marker values for cases and controls when event rate is 0.05. We expect to use this plot to explain the similarity of AUC_{t_0} and diversity of AP_{t_0}. It seems that AUC_{t_0} is associated with the overlapping area of the two density curves for cases and controls. In these two plots of Figure 3.1, the overlapping areas appear to be similar, and correspondingly the two AUC_{t_0} values are similar. On the other hand, AP_{t_0} seems to be associated with the right tail area of the cases' density curve beyond the density curve of controls. The larger this area is, the larger the value of AP_{t_0} appears. The right tail area in the right plot seems larger compared to the area in the left plot. Correspondingly, the value of AP_{t_0} for the right plot is larger than the one for the left plot. Further theoretical verification is required.

This simulation has shown that AP_{t_0} is an attractive alternative to AUC_{t_0} for comparing markers. Especially when the markers have the same AUC_{t_0} and our goal is to detect high risk groups, we can use AP_{t_0} to measure the risk prediction models. In addition, AP_{t_0} is sensitive to event rate, which provides more information.

Chapter 4

Data Analysis

Here we use an example to illustrate our proposed time-dependent AP on the Childhood Cancer Survivor Study (CCSS) data [23]. The dataset consists of 13060 observations with survival data, of which 11437 subjects who have complete information are analyzed. 98.7% of the analyzed subjects are censored, while the remaining 248 individuals who developed significant cardiovascular disease are observed. In the study by Chow et al. [23], the subjects are the survivors who were free of significant cardiovascular disease 5 years after cancer diagnosis from the CCSS cohort. The children in the study were diagnosed with cancer before age 21 and were enrolled in 26 institutions in North America between 1970 and 1986. The follow-up stopped when the participants turned 40 years old. The goal of Chow et al. [23] was to create clinically useful risk groups that incorporate demographic and cancer treatment information available at the end of therapy to predict subsequent congestive heart failure (CHF) among participants. Even though CHF risk predictors can be found in the general older adult population, cardiovascular disease is becoming increasingly recognized as one of the top contributors to late morbidity and mortality in the now more than 400000 childhood cancer survivors in the United States. Seeing that validated CHF risk prediction models specified for adolescent and young adult survivors were not available, they hoped that a robust CHF prediction model can be created to help clinicians better identify and counsel this population with higher risk of CHF.

The data analyzed in our study includes nine risk score systems developed by Chow et al. [23]: `simple_riskscore, simple_riskgroup, heart_riskscore, standard_riskgroup, heart_riskgroup, standard_logP, standard_riskscore, simple_logP, heart_logP`. 'Simple', 'heart' and 'standard' represent three different risk prediction models: 'simple' refers to a simple model where chemotherapy and radiotherapy treatment were categorized as yes or no, 'standard' is a standard model where clinical dose information was known, and 'heart' is a standard with heart dose model which uses average radiation dose to the heart in lieu of chest field dose. Meanwhile, 'logP', 'riskgroup', and 'riskscore' are three risk subsystems. 'Riskscore' stands for the integer risk scores converted from the predicted relative risks through Poisson regression models. 'Riskgroup' represents three risk groups low, moderate, and high, which are created based on 'riskscore'. 'LogP' is a linear predictor from the regression model.

We estimate AUC_{t_0} and AP_{t_0} for the nine risk scores developed by Chow et al. [23] with different prediction time points t_0 ranging from year 5 to year 35. These estimates are shown in Figure 4.1 and Figure 4.2. The numeric results and the inferences are provided in Table C.1 and C.2 from Appendix C.

Figure 4.1 shows that: (1) among the three risk prediction models, 'logP' has higher AUC_{t_0} than 'riskscore' in general, and 'riskgroup' has the lowest AUC_{t_0} within each model, (2) comparing the three prediction models, 'heart' and 'standard' are indistinguishable for the first two years and last five years with 'logP' and 'riskscore', but 'heart' model seems to be better than 'standard' between the two time intervals, (3) 'heart' overwhelms both 'standard' and 'simple' with risk subsystem 'riskgroup', (4) and among the nine risk score systems, 'heart_logP' and 'standard_logP' have better performances than any others. 'Heart_logP' is an even better choice than 'standard_logP' from year 7 to year 30 based on AUC_{t_0}-scale. The line at the bottom of this figure is a reference line which indicates that AUC_{t_0} is equal to 0.5 for a random classifier. Theoretically, all useful risk score systems should be above this line.

Figure 4.2 shows the performances of the nine risk score systems in terms of AP_{t_0}. The results show that: (1) within each risk subsystem ('logP', 'riskscore', 'riskgroup'), 'heart' outperforms 'simple' and 'standard', (2) within the prediction model 'heart', 'riskgroup' has the most competitive performance, while 'logP' outperforms 'riskscore'. With prediction models 'simple' and 'standard', 'logP' has a slightly better performance than 'riskscore', and 'riskscore' overwhelms 'riskgroup', (3) comparing these nine risk scores, 'heart_riskgroup' outperforms every other score. The line at the bottom of the plot indicates the event rate over time, which is used as a reference line. Theoretically, all useful risk score systems should be above this line.

Comparing Figure 4.1 and 4.2, we find that some riskscore systems have similar AUC_{t_0} but different AP_{t_0}. For example, 'heart_logP' and 'standard_logP' share similar AUC_{t_0} but different AP_{t_0} over time, which can be shown by the upper two plots in Figure 4.3. On the contrary, some riskscore systems share similar AP_{t_0} but different AUC_{t_0}. This situation can be shown by the lower two plots of riskscore systems 'standard_riskscore' and 'standard_logP' in Figure 4.3.

Depending on the different performance measures AUC_{t_0} and AP_{t_0}, we arrive at different conclusions for the choice of risk score system. Based on the AUC_{t_0}-scale plot, 'heart_logP' and 'final_logP' are preferred, whereas the AP_{t_0}-scale plot favours the 'heart_riskgroup', which has a relatively small AUC_{t_0}. However, both measures suggest that 'heart' is the best prediction model over time. In addition, when two riskscore systems share similar performance over time based on one of the two measures, they might have different performances depending on the other measure. Chow et al. [23] used AUC and C-statistics as the main performance measures, thus the proposed assessment provided by AP_{t_0} as discussed might provide them with delightful insights into their study.

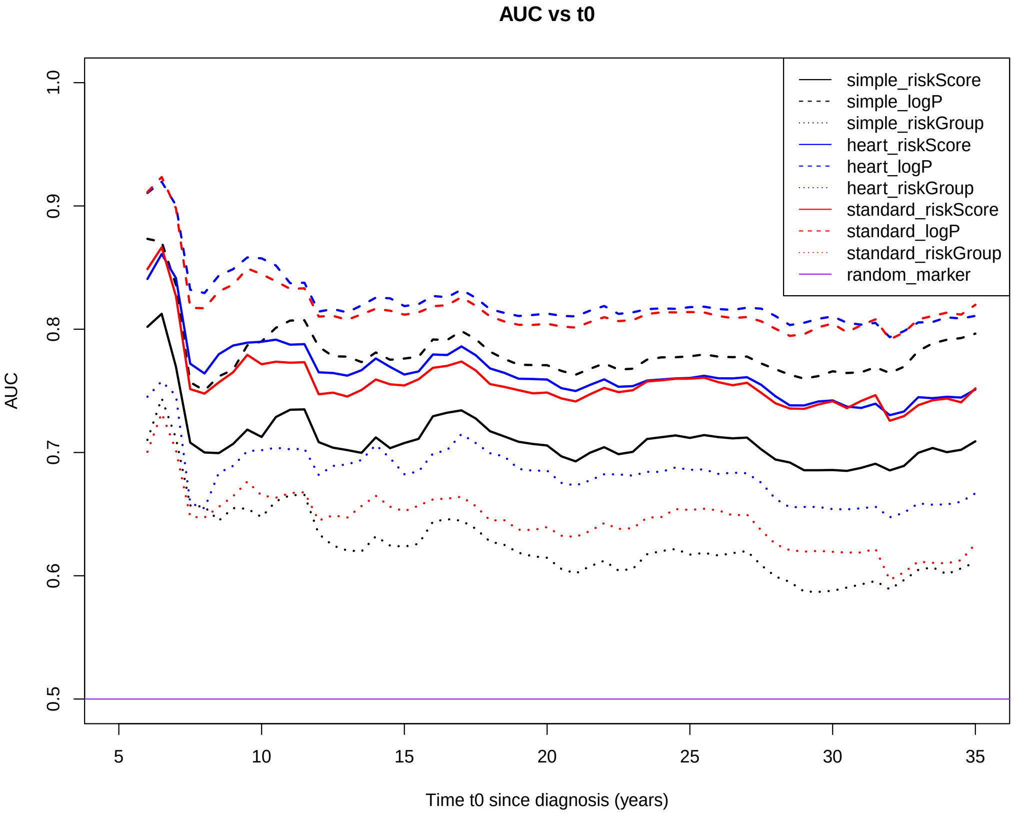

Figure 4.1: Risk Score System's Time-dependent AUC.

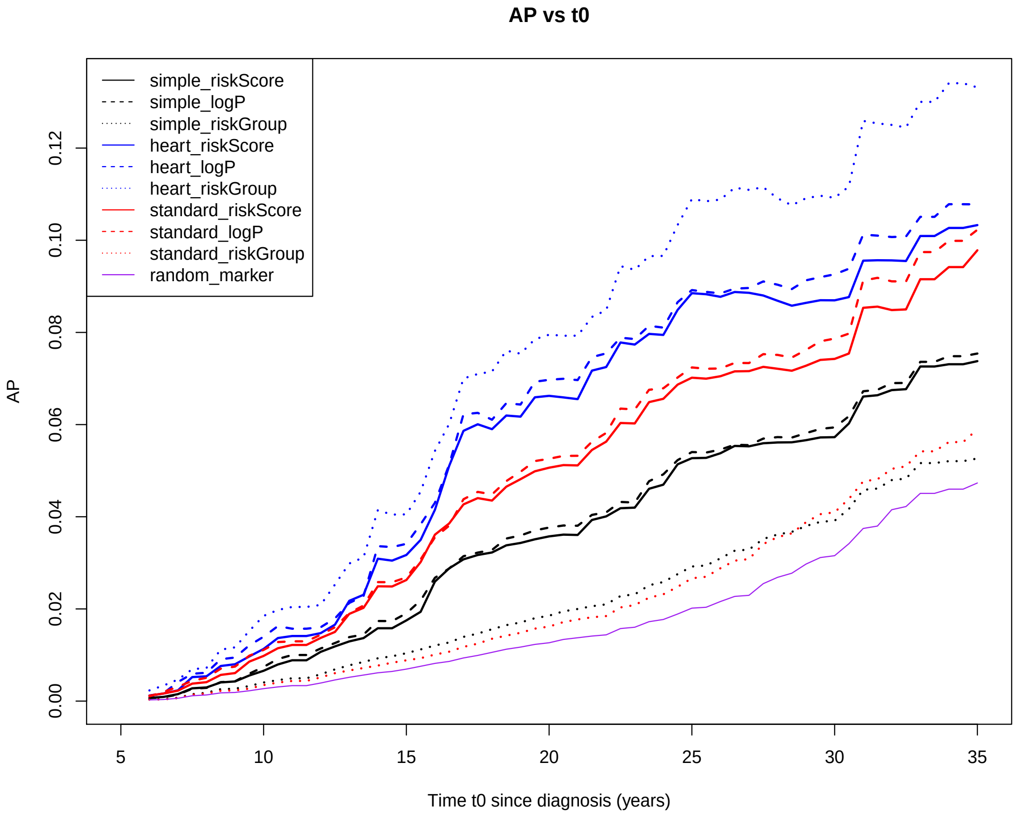

Figure 4.2: Risk Score System's Time-dependent AP.

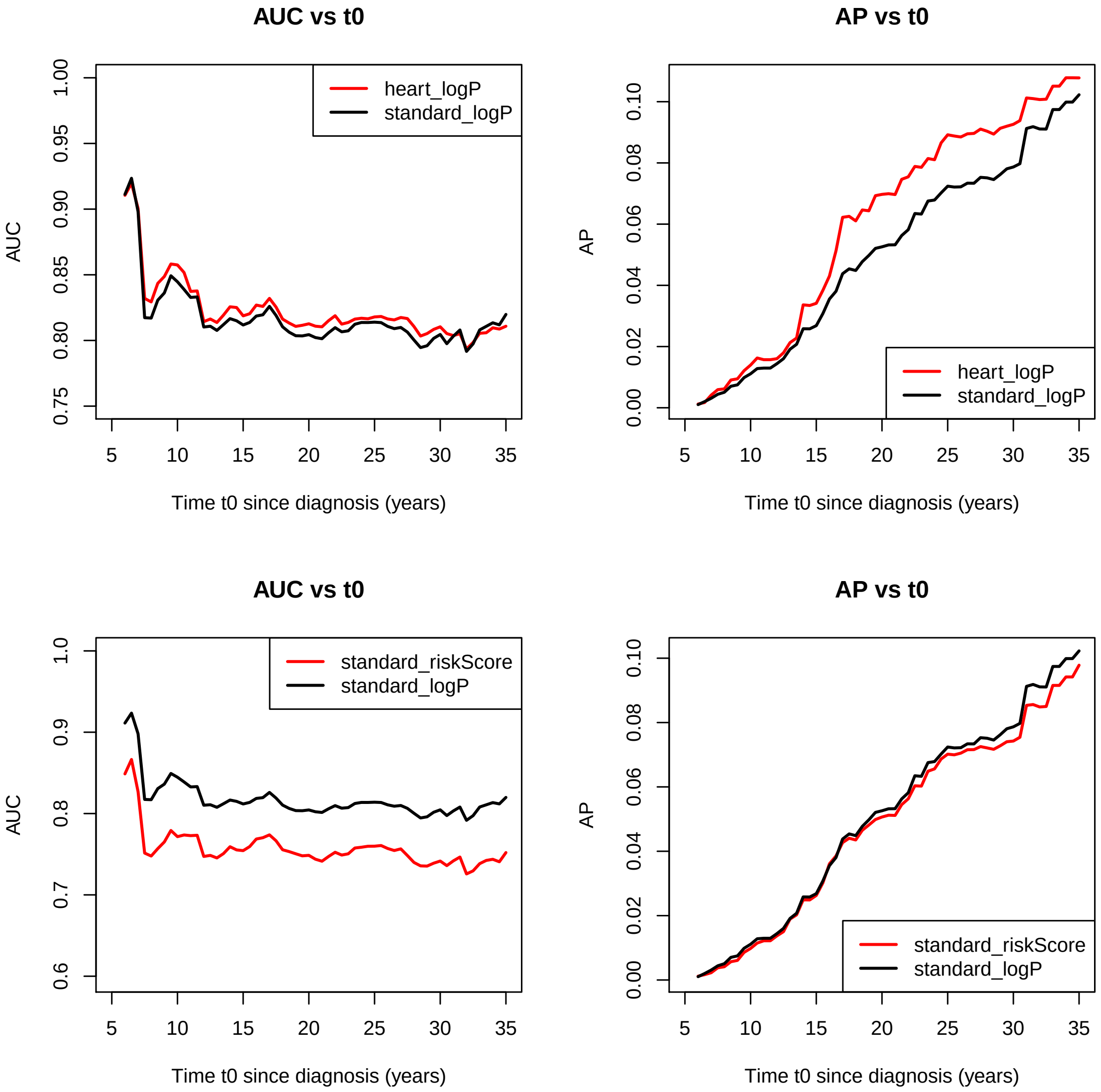

Figure 4.3: Two Pairs Riskscore Systems' AUC$_{t_0}$ & AP$_{t_0}$ Comparisons

Chapter 5

Discussion

5.1 Summary

In this project, we extend the AP, which was proposed by Yuan et al. [1] for binary outcomes, to accommodate time-dependent outcomes. Nonparametric estimates of the proposed measure are derived. In the presence of censoring, for some subjects, their time-dependent disease outcomes are unknown. Thus, the IPW technique is applied to deal with the missing disease outcomes due to censoring. Then bootstrap and perturbation resampling are applied to construct confidence intervals. We conducted simulation studies to examine the performance of estimated AUC_{t_0} and AP_{t_0} through finite samples. The results show that the performance of the estimated time-dependent AUC and AP is satisfactory with ignorable biases and accurately estimated standard errors. We need to point out that by the design of the simulation studies, the estimated values of AUC_{t_0} are similar in each scenario and the estimates of AP_{t_0} are different in the two scenarios. This can help us to understand how distributions of marker values for cases and controls is related to time-dependent AUC and AP. More simulation studies might be required for further investigation.

We realize that the estimation of time-dependent AP is not stable when the number of cases is small, since calculation of the AP depends on the number of cases. If the event rate is small, we need a large sample size to have satisfactory performance of the estimation and inference procedures.

Lastly, we estimated the time-dependent AUC and AP, and evaluated the predictive performance of the nine risk score systems in a real-life dataset — CCSS data. Depending on different measurements: AUC_{t_0} and AP_{t_0}, we arrive at different choices of the risk score system, but a certain unique prediction model is favoured.

5.2 Future Investigation

In this project, we propose the time-dependent AP for a single marker. In reality, clinicians construct risk prediction models with more than one significant markers. The t_0-year risk can be estimated parametrically or nonparametrically. We intend to extend the time-dependent AP to accommodate multiple markers. Parametric and nonparametric estimates of the time-dependent AP can be developed.

In practice, clinicians are also interested in investigating whether adding new markers on top of the existing markers could improve the risk predictive performance. Differences in AP can also be used as a measure to quantify the incremental value of the new markers by comparing the performance of the new prediction model with both existing markers and new markers with the old prediction model.

In some applications, clinicians usually want to measure the predictive performance when FPF is below a certain threshold, such as 5% or 10%. Therefore, partial AP (area under a specific range of the precision-recall curve) might also be of interest when they want to apply AP to evaluate prediction models given a prespecified range of FPF.

Bibliography

[1] Yan Yuan, Wanhua Su, and Mu Zhu. Threshold-free measures for assessing the performance of medical screening tests. *Frontiers in Public Health*, 3, 2015.

[2] D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics.* New York: Wiley, 1966.

[3] James P. Egan. *Signal detection theory and ROC-analysis.* Academic Press New York, 1975. ISBN 0122328507.

[4] John A Hanley et al. Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging*, 29(3):307–35, 1989.

[5] Margaret Sullivan Pepe. A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, 84(3):595–608, 1997.

[6] Margaret Sullivan Pepe. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, 56(2):352–359, 2000.

[7] Charles E Metz. ROC methodology in radiologic imaging. *Investigative Radiology*, 21 (9):720–733, 1986.

[8] Tianxi Cai and Margaret Sullivan Pepe. Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American statistical Association*, 97(460):1099–1107, 2002.

[9] Kelly H Zou, WJ Hall, and David E Shapiro. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 16 (19):2143–2156, 1997.

[10] Charles E Metz, Benjamin A Herman, and Jong-Her Shen. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, 17(9):1033–1053, 1998.

[11] Donna Katzman McClish. Analyzing a portion of the ROC curve. *Medical Decision Making*, 9(3):190–195, 1989.

[12] Mary Lou Thompson and Walter Zucchini. On the statistical analysis of ROC curves. *Statistics in Medicine*, 8(10):1277–1290, 1989.

[13] Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, pages 337–344, 2000.

[14] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[15] SC Cheng, LJ Wei, and Z Ying. Analysis of transformation models with censored data. *Biometrika*, 82(4):835–845, 1995.

[16] Hajime Uno, Tianxi Cai, Lu Tian, and LJ Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478), 2007.

[17] Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, 2005.

[18] Chaya S Moskowitz and Margaret S Pepe. Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics*, 5(1): 113–127, 2004.

[19] Yingye Zheng, Tianxi Cai, Margaret S Pepe, and Wayne C Levy. Time-dependent predictive values of prognostic biomarkers with failure time outcome. *Journal of the American Statistical Association*, 103(481):362–368, 2008.

[20] Yingye Zheng, Tianxi Cai, Janet L Stanford, and Ziding Feng. Semiparametric models of time-dependent predictive values of prognostic biomarkers. *Biometrics*, 66(1):50–60, 2010.

[21] Chien-Fu Jeff Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, pages 1261–1295, 1986.

[22] Enno Mammen. Bootstrap, wild bootstrap, and asymptotic normality. *Probability Theory and Related Fields*, 93(4):439–455, 1992.

[23] Eric J Chow, Yan Chen, Leontien C Kremer, Norman E Breslow, Melissa M Hudson, Gregory T Armstrong, William L Border, Elizabeth AM Feijen, Daniel M Green, Lillian R Meacham, et al. Individual prediction of heart failure among childhood cancer survivors. *Journal of Clinical Oncology*, 33(5):394–402, 2015.

Appendix A

The probability expression of AP

Let $S_1(c) = P(Z_{1i} > c)$ denote a survival function of cases, where Z_{1i} is the marker value of cases. $F_1(c) = P(Z_{1i} \leq c)$ denotes the cumulative density function of cases and $f_1(c)$ is the corresponding probability density function.

It is known that

$$AP = \int_{\infty}^{-\infty} PPV(c) dTPF(c) = \int_{\infty}^{-\infty} \frac{P(Z_i > c, D_i = 1)}{P(Z_i > c)} d\, S_1(c)$$

$$= \int_{\infty}^{-\infty} \frac{P(Z_i > c \mid D_i = 1)P(D_i = 1)}{P(Z_i > c \mid D_i = 1)P(D_i = 1) + P(Z_i > c \mid D_i = 0)P(D_i = 0)} (-f_1(c)) dc$$

$$= \int_{-\infty}^{\infty} \frac{P(Z_i > c \mid D_i = 1)P(D_i = 1)}{P(Z_i > c \mid D_i = 1)P(D_i = 1) + P(Z_i > c \mid D_i = 0)P(D_i = 0)} f_1(c) dc$$

Assume that $P(D_i = 1) = \pi$, which is the event rate, and $S_0(c) = P(Z_{0i} > c)$ denote the survival function for the control. Then,

$$\begin{aligned} AP &= \int_{-\infty}^{\infty} \frac{S_1(c)\pi}{S_1(c)\pi + S_0(c)(1-\pi)} f_1(c) dc \\ &= E\left\{ \frac{S_1(Z_{1j})\pi}{S_1(Z_{1j})\pi + S_0(Z_{1j})(1-\pi)} \right\} \\ &= E\left[P(D_i = 1 \mid Z_i > Z_{1j}) \right]. \end{aligned}$$

Appendix B

The transformed estimates' coverage probability in simulation

We summarized the original and transformed coverage probability of time-dependent AUC and AP estimates in Chapter 3 with two tables: one is for bootstrap and the other one is for perturbation resampling results.

		Sample Size 5000		Sample Size 10000	
		Original ECOVP	Transformed ECOVP	Original ECOVP	Transformed ECOVP
r=0.01	AP N(0.95,1)	0.857	0.906	0.903	0.921
	AP N(1.51,2)	0.931	0.974	0.947	0.961
r=0.05	AP N(0.95,1)	0.926	0.931	0.936	0.936
	AP N(1.51,2)	0.934	0.938	0.952	0.954
r=0.1	AP N(0.95,1)	0.946	0.949	0.957	0.959
	AP N(1.51,2)	0.952	0.954	0.95	0.95

Table B.1: Transformed AP Estimates Coverage Probability for Bootstrap.

		Sample Size 5000		Sample Size 10000	
		Original ECOVP	Transformed ECOVP	Original ECOVP	Transformed ECOVP
r=0.01	AP N(0.95,1)	0.853	0.883	0.898	0.911
	AP N(1.51,2)	0.919	0.952	0.94	0.955
r=0.05	AP N(0.95,1)	0.917	0.92	0.929	0.933
	AP N(1.51,2)	0.932	0.939	0.951	0.954
r=0.1	AP N(0.95,1)	0.941	0.944	0.953	0.956
	AP N(1.51,2)	0.95	0.953	0.948	0.948

Table B.2: Transformed AP Estimates Coverage Probability for Perturbation Resampling.

Appendix C

Numeric Results of CCSS dataset

Two tables below represent the numeric estimates and inferences of the nine markers from CCSS data with specified t_0 and the corresponding event rate r. ‘Estimate’ stands for the point estimate and ‘SE’ is the standard error obtained via perturbation resampling.

	$t_0 = 10$, $r = 0.03$	$t_0 = 15$, $r = 0.07$	$t_0 = 20$, $r = 0.013$	$t_0 = 25$, $r = 0.02$	$t_0 = 30$, $r = 0.032$	$t_0 = 35$, $r = 0.047$
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
simple_riskScore	0.713 (0.04)	0.708 (0.025)	0.706 (0.022)	0.712 (0.018)	0.686 (0.017)	0.709 (0.019)
simple_logP	0.79 (0.036)	0.776 (0.023)	0.771 (0.02)	0.778 (0.017)	0.766 (0.016)	0.796 (0.019)
simple_riskGroup	0.648 (0.04)	0.624 (0.026)	0.615 (0.023)	0.617 (0.019)	0.588 (0.017)	0.611 (0.021)
heart_riskScore	0.79 (0.037)	0.763 (0.025)	0.759 (0.021)	0.76 (0.017)	0.742 (0.016)	0.751 (0.019)
heart_logP	0.857 (0.034)	0.819 (0.023)	0.813 (0.02)	0.818 (0.016)	0.81 (0.015)	0.811 (0.018)
heart_riskGroup	0.702 (0.04)	0.682 (0.029)	0.685 (0.024)	0.686 (0.019)	0.654 (0.018)	0.667 (0.023)
standard_riskScore	0.772 (0.038)	0.754 (0.026)	0.749 (0.022)	0.76 (0.017)	0.742 (0.017)	0.752 (0.022)
standard_logP	0.845 (0.032)	0.812 (0.023)	0.804 (0.02)	0.814 (0.016)	0.805 (0.015)	0.82 (0.019)
standard_riskGroup	0.665 (0.038)	0.652 (0.028)	0.639 (0.023)	0.653 (0.019)	0.619 (0.019)	0.626 (0.026)

Table C.1: Numeric Results of AUC for CCSS.

	$t_0 = 10$, $r = 0.03$	$t_0 = 15$, $r = 0.07$	$t_0 = 20$, $r = 0.013$	$t_0 = 25$, $r = 0.02$	$t_0 = 30$, $r = 0.032$	$t_0 = 35$, $r = 0.047$
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
simple__riskScore	0.007 (0.001)	0.018 (0.003)	0.036 (0.006)	0.053 (0.008)	0.057 (0.006)	0.074 (0.008)
simple__logP	0.007 (0.002)	0.019 (0.003)	0.038 (0.005)	0.054 (0.007)	0.059 (0.006)	0.075 (0.007)
simple__riskGroup	0.004 (0.001))	0.010 (0.001)	0.019 (0.002)	0.029 (0.002)	0.039 (0.003)	0.053 (0.005)
heart__riskScore	0.011 (0.003)	0.032 (0.008)	0.066 (0.013)	0.089 (0.015)	0.087 (0.011)	0.103 (0.012)
heart__logP	0.014 (0.004)	0.034 (0.009)	0.070 (0.015)	0.089 (0.014)	0.093 (0.012)	0.108 (0.012)
heart__riskGroup	0.018 (0.006)	0.041 (0.009)	0.080 (0.014)	0.109 (0.016)	0.109 (0.015)	0.133 (0.019)
standard__riskScore	0.010 (0.003)	0.026 (0.006)	0.051 (0.009)	0.070 (0.010)	0.074 (0.008)	0.098 (0.012)
standard__logP	0.011 (0.003)	0.027 (0.005)	0.053 (0.009)	0.072 (0.010)	0.079 (0.008)	0.102 (0.012)
standard__riskGroup	0.003 (0.001)	0.009 (0.001)	0.016 (0.002)	0.027 (0.002)	0.041 (0.003)	0.059 (0.005)

Table C.2: Numeric Results of AP for CCSS.

Appendix D

ROC and precision-recall curves for CCSS data

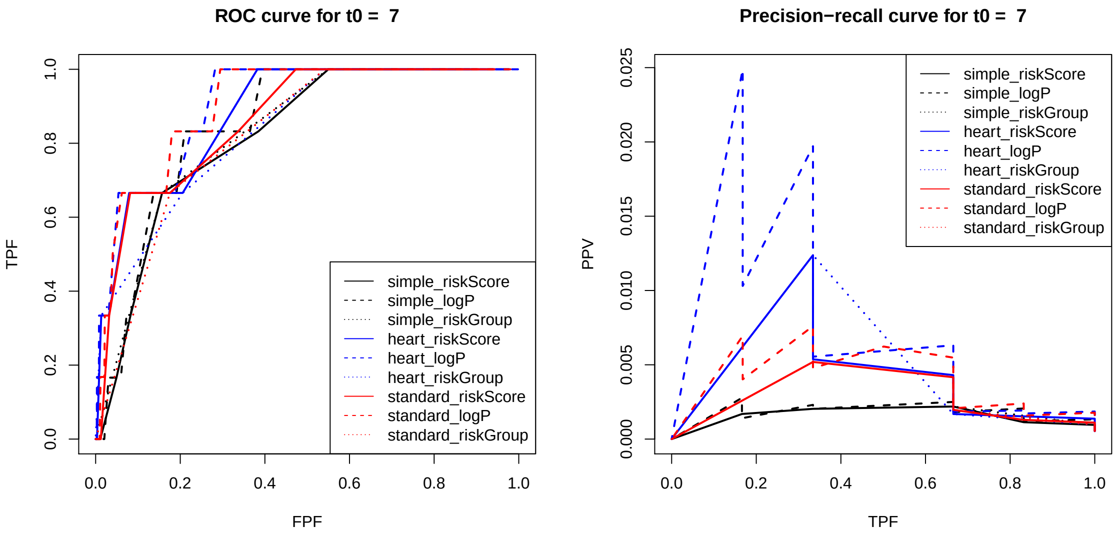

Figure D.1: ROC and PR curve : t_0=7 (year).

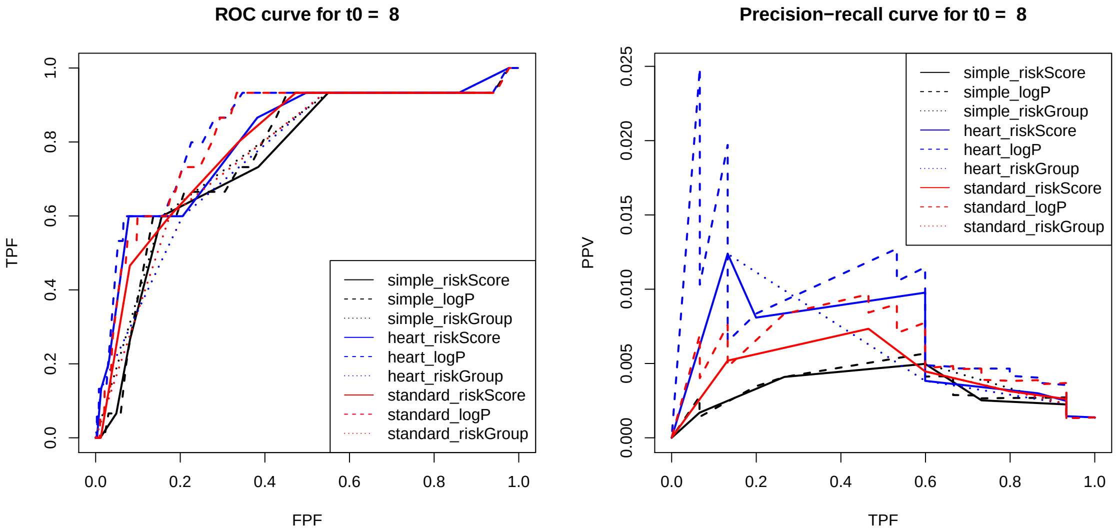

Figure D.2: ROC and PR curve : t_0=8 (year).

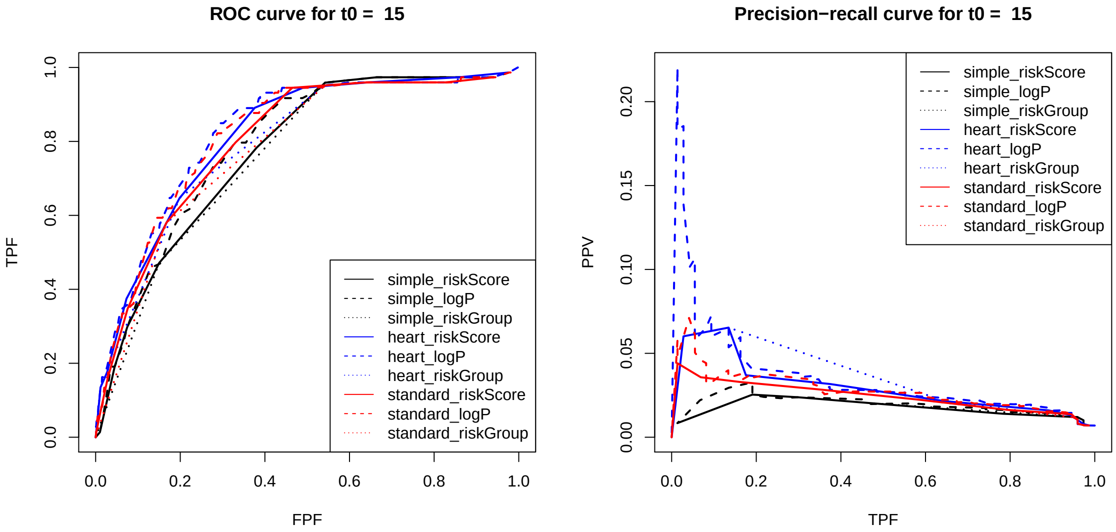

Figure D.3: ROC and PR curve : t_0=15 (year).

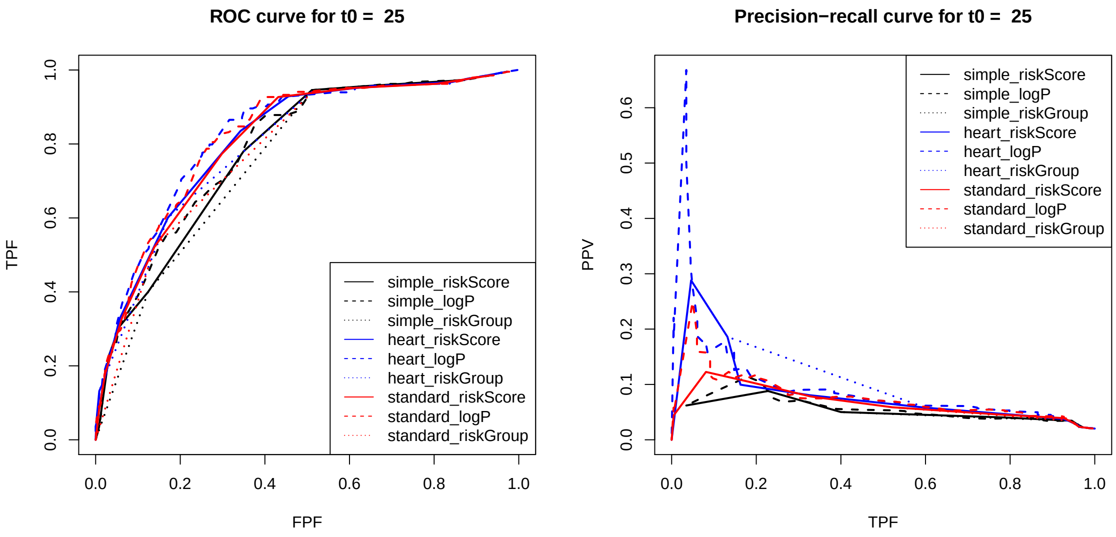

Figure D.4: ROC and PR curve : t_0=25 (year).

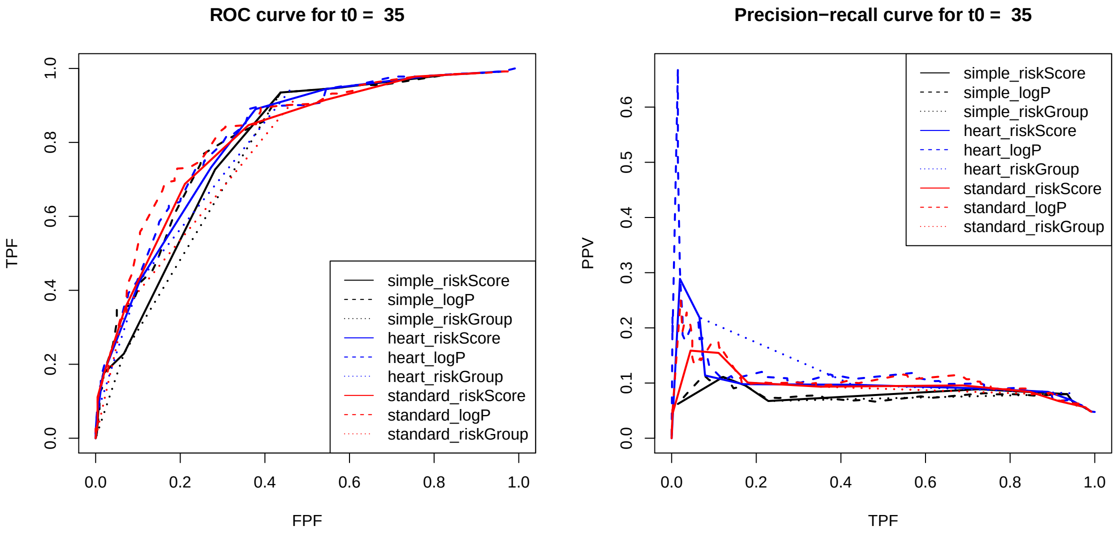

Figure D.5: ROC and PR curve : t_0=35 (year).